The measurement of sickness impact : construction of the SIP68

The measurement of sickness impact

Construction of the SIP68

Proefschrift

ter verkrijging van de graad van doctor aan de Rijksuniversiteit Limburg te Maastricht, op gezag van de Rector Magnificus, Prof. Mr. M.J. Cohen, volgens het besluit van het College van Dekanen, in het openbaar te verdedigen

op vrijdag 12 april 1996 om 16.00 uur

door

Ate Frans de Bruin

geboren 27 juli 1960 te Groningen

promotor
prof. dr. H. Philipsen

co-promotor
dr. J.P.M. Diederiks

beoordelingscommissie
prof. dr. J.A. Knottnerus (voorzitter)
prof. dr. A.P.W.M. Appels
prof. dr. G.A.M. van de Bos, (Universiteit van Amsterdam)
dr. B. Tax, (Katholieke Universiteit Nijmegen)
prof. dr. F. Sturmans

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van:

- Researchfonds, Ziekenhuis De Wever & Gregorius te Heerlen/Brunssum
- IRV te Hoensbroek
- Defauwes, orthopedische schoentechniek
- Hanssen, orthopedische schoentechniek
- Smeets & zn., orthopedische schoentechniek

Contents

'Ten eerste', zei Hans Castorp, die zijn woorden nu zelfs al indeelde in ten eerste en ten tweede, 'begrijp ik niet, waarom je met een onschuldige koorts - en dan wil ik voorlopig even aannemen, dat zoiets bestaat - met onschuldige koorts het bed zou moeten houden maar met een andere niet. En ten tweede zeg ik je toch dat ik me sinds die verkoudheid niet verhitter voel dan tevoren al. Ik sta op het standpunt', besloot hij, 'dat 37,6 gelijk is aan 37,6. Als jullie daarmee kunnen rondlopen, kan ik het ook.'

(De Toverberg, Thomas Mann, vertaling Pé Hawinkels)

Voor mijn ouders

General introduction

The WHO definition of health 'the state of optimal physical psychological and social well-being' lead to an increased attention to a broad view on health. No longer only bio physiological data was considered to be an adequate indicator of health status, data on the subjective experience of health, and behavior or functioning in relation to health, also became relevant information in health research. This development was stimulated by the increase in chronic or incurable health deviations. To be able to study these non-biophysiologic aspects of health, instruments were constructed that aimed to assess these concepts. Many of these instruments were questionnaires or observation checklists. Gradually this kind of instruments became accepted as relevant sources of information on health status.
One of the most widely used questionnaires of this kind is the Sickness Impact Profile (SIP). The SIP is a generic questionnaire assessing the influence of health deviations on functioning or the health related functional status. The instrument was developed and introduced in the USA by Bergner and her group in the late seventies. During the eighty's it was translated and used in numerous health research projects in several European countries. It gained the reputation of a valid and reliable measure of functional status. Hence, the instrument was a core source of information in numerous research projects both in the USA and in Europe.
A major drawback of the SIP, however, is its length, it takes 20 to 30 minutes to respond to all 136 questions, while other instruments measuring related concepts take conciderably less time to complete.
In 1989 the SIP-project started as a joint project at the University of Limburg in Maastricht and at the IRV in Hoensbroek. The aim of this project was to evaluate and, if found necessary, modify or shorten the SIP without losing the generic character or the high psychometric quality. The year the project started, several attempts were made to contact prof. Bergner in order to inform her on our work on the SIP and to ask her comment. No reaction was recieved.

This thesis is a report on the SIP-project which ended in 1995. Chapter 1 is the core of the thesis, and the quick reader may confine himself to this part. After a short introduction of the SIP (paragraph 1.3), a model of health related behavior is developed (paragraph 1.4). This model is meant to offer a frame of reference to clarify the concept of health related functional status, and to support the interpretation of data concerning this concept. The implications of the health behavior model for data provided by the SIP are discussed in subparagraph 1.3.6. The other paragraphs in chapter 1 provide a report on the rest of the SIP project from an 'eagle's view'. Paragraph 1.4 provides a global description of the findings of a literature review on the psychometric characteristics of the SIP. The next paragraph (1.5) describes how a short version of the SIP the SIP68, was constructed.

The testing of the reliability and validity of this short instrument is described in paragraph 1.6. Paragraph 1.7 describes methods and findings of a study into the responsiveness of the original SIP and the SIP68. A general discussion of the SIP project is presented in paragraph 1.8.
The chapters 2 to 7 describe the aspects of the project described in paragraph 1.4 to 1.7 in more detail. Chapter 2 provides an extensive description of the methods and findings of a literature review into the psychometric characteristics of the original Sickness Impact Profile. The methods used to develop the SIP68 are elaboratly preseted in chapter 3. Chapter 4 and 5 describe two successive projects in which the reliability and validity of the SIP68 are studied in a rheumatic population and in a spinal cord injured population respectively. The next chapter (6) describes a study into the level to which both the original SIP and the SIP68 are able to detect changes, the responsiveness of both instruments. Finally, chapter 7 presents results on analyses performed on the items that were dropped during the construction of the short SIP68. The question addressed to in this chapter is: did we lose essential information on the health related functional status by dropping these items.

Chapter 1
The SIP-project

1.1. Introduction

In the study and evaluation of health and health care during the last decades, the use of questionnaires to measure (aspects of) health status has been seen to increase. This development is generally attributed to the growing consensus that measures of experienced health or functional status should be integrated in the study design. Although a simplification, two types of general health questionnaires can be distinguished: measures of functional status and measures of experienced health. The latter instruments ask the respondents to indicate how they feel concerning different aspects of health or health in general. They can be considered measures of well-being. The former instruments assess health by measuring deviations in functioning or behavior resulting from health deviations. One of the best known and most widely used questionnaires of the functional status type is the Sickness Impact Profile (SIP)[1]. The SIP is meant to be a generic measure of health related functional status. It was developed in the late 1970s in the USA and has been used in many studies in different populations, both in America and in Europe. In 1985 the SIP was translated into Dutch [2] and used in research projects. At the University of Limburg in Maastricht and the Institute for Rehabilitation Research (IRV) in Hoensbroek (both in the Netherlands) a number of projects used the SIP as central source of information. Thus, a database containing over 2500 completed SIP lists from ten different diagnostic groups was available for secondary analysis.

When a researcher wants to select the most suitable measure of functional status from the large number of instruments available, two general problems occur. First, most functional status instruments lack a theoretical frame of reference that justifies the way they are developed and that might guide the interpretation of findings. This problem concerns the construct validity of functional status measures. Secondly, it is not known to what degree most instruments are able adequately to register relevant changes in the concept measured. A problem specific for the SIP is its length. The SIP contains 136 items and on this point compares unfavorably with other functional status measures. Hence, it was decided to start a project aimed at evaluating and adapting the SIP. Three topics were incorporated in this SIP project: the theoretical foundation of functional status measures, especially the SIP, the psychometric characteristics and responsivenss of the SIP, and the development of a short version of the instrument. This thesis is a report on that SIP project.
After a short introduction to the list, the theory behind the concept of 'health-related functional status' will be explored. Theoretical literature will be used to

construct a frame of reference for the instrument. Next, as the SIP is generally known to be a reliable and valid measure, but evidence for this claim is only found in incidental studies that were not directly aimed at evaluating the SIP, a review of methodological literature will be performed to study the psychometric properties of the list. After this, the SIP database will be used to develop a short generic version of the instrument. Finally, the longitudinal data available will be used to study the ability of the instrument to detect changes in the health-related functional status (responsiveness).

1.2. The SIP

The Sickness Impact Profile consists of 136 items, every item being a statement on behavior. The items are grouped into twelve categories, every category covering an aspect of daily functioning. Two dimensions are distinguished: a physical dimension consisting of three categories and a psychosocial dimension composed of four categories. The other categories are not aggregated [1] (see figure 1.1). Respondents are asked to check those items that describe their situation on the day they fill out the list, but do so only when the fact that these items apply to them is related to their health status. Scores for every category, for both dimensions and for the overall instrument are calculated after attaching differential weights to every item. Scores range from 0 to 100, with higher scores representing more dysfunction [1].

Many researchers have used the instrument and based their conclusions on it, but up to 1989 no systematic evaluation of the SIP as a measure of general health had been published. In the SIP project, international methodological literature was systematically reviewed to present the state of the art of the SIP [3]. Findings are presented in paragraph 1.4. First, however, the concept measured by the SIP 'health-related functional status' will be discussed in more detail.

1.3. The concept 'Health related behavorial changes'

The SIP is a general health measure that operationalizes health in functional or behavioral terms. It registers behavioral or functional limitations resulting from health deviations. In general functional limitations might result from pathology, but not necessarily so. A person might experience his health as deviant without any objectively detectable pathology. The same occurs in behavior changes: this might result from a pathological process, but also a social or psychological process might underlie these behavioral changes. In filling out the SIP respondents are asked to decide whether or not certain behavioral changes are attributable to health status. These health- related behavioral changes are used as an indicator for health status. Although in many research studies functional status measures like the SIP are central sources of information, systematic theorizing about this 'translation

Figure 1.1. Categories and item examples of the SIP

Independent categories

SR Sleep and Rest (7 items)
- I sleep or nap during the day.

E Eating (9 items)
- I am eating special or different food.

W Work (10 items)
- I often act irritable toward my work associates.

HM Home Management (10 items)
- I am not doing heavy work around the house.

RP Recreation and Pastimes (8 items)
- I am going out for entertainment less.

Physical Dimension

A Ambulation (12 items)
- I walk shorter distances or stop to rest often.

M Mobility (10 items)
- I stay within one room.

BCM Body Care and Movement (23 items)
- I am very clumsy in body movements.

Psychosocial Dimension

SI Social Interaction (20 items)
- I am doing fewer social activities with groups of people.

AB Alertness Behavior (10 items)
- I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things.

EB Emotional Behavior (9 items)
- I laugh or cry suddenly.

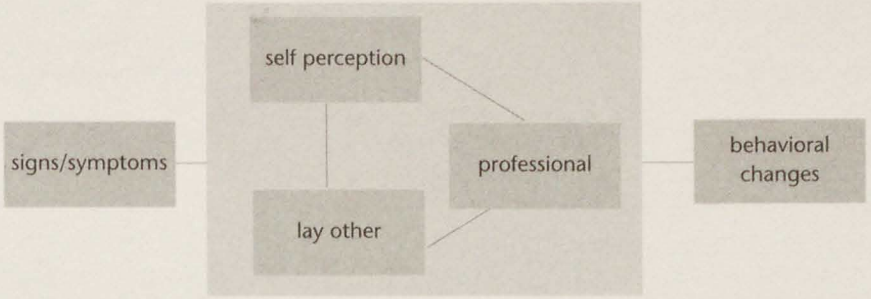
C Communication (9 items)
- I am having trouble writing or typing.

between behavioral changes and health status' is rarely found. Hence no theoretical frame of reference is present to guide interpretation of findings of this type of instrument. On a fundamental level, therefore, a model of health-related behavioral consequences will be developed to clarify the concept of health-related functional status. Conceptual clarity will enable researchers to make a more thoroughly motivated choice whether to use this kind of instrument to answer a given research question.

1.3.1. A model of behavioral health consequences

A model of how consequences of health deviations are generated starts with the disturbance of the normal healthy situation. A person feels something or notices that he/she cannot perform activities that he or she is used to doing or considers normal to do or to be able to do. It might also be that it is not the person himself but someone else who notices a change or something strange in the behavior of the individual in question, e.g. memory problems or hearing loss in ageing persons. As with any other experience, the individual will try to interpret this information and will need a frame of reference. Are these observations 'true' or 'real' and what causes can be identified? After lay consultation in the person's social environment, an expert (e.g. physician) might be consulted to verify things and possibly to do something about it. The person in question now has become a patient and the physician may legitimize the situation by means of a medical diagnosis. The former diffuse meanings and feelings are now placed within a medical frame of reference. Experienced changes are identified as consequences of health deviations and causes are codified in a diagnosis. However firm this diagnosis may be, hardly any officially codified disturbance of health is connected with a clear-cut pattern of behavioral changes. Hence, in these cases there is much room for interpretation. All health care professionals know that the behavioral consequences within a given diagnosis may vary strongly between individuals. This explains why physicians often are reluctant to give definitive answers to the question of what concrete changes are to be expected in an individual case. However, to maintain an understandable, meaningful and predictable daily life, the individual in question needs to know what he can expect from his family, social environment and the health professionals he encounters. A person experiencing health deviations and his social environment, therefore, will define a set of consequences to be expected, given the perceived health status of the person in question. Given the loose definition of the behavioral consequences of health deviations, the individual set of health-related behavioral consequences will result from implicit or explicit beliefs, communications and negotiations between the individual in question and significant others (lay and health professionals). A new set of role expectancies has to be defined, based on the evaluation of the new health situation. This line of thinking led to the development of figure 1.2.

Figure 1.2. Model of the connection between signs/symptoms and health related behavioral changes.



The model reads as follows: someone experiences a change in his situation. These changes are communicated with lay and professional others. It might also be that others notice these changes and communicate them to the individual in question. All three parties reflect on the situation and a certain level and type of health-related behavioral changes occurs. Every change in the health status due to treatment or a shift in the underlying medical situation might start the process of negotiating on the expected set of behavioral consequences anew. Hence the model in figure 1.2 has to be viewed as an ongoing and circular process of feedback and adjustment.

The theoretical tenability of this model will now be elaborated upon. Then the implications of this model as frame of reference for the interpretation of information on the health-related functional status and for health care will be explored.

1.3.2. Health and behavior

One of the first and most elaborated theories on the relation between health and consequent behavior is found in Parsons' work on the sick role. Parsons defines health as 'the state of optimum capacity of an individual for the effective performance of the roles and tasks for which he has been socialized' [4]. This health definition refers to roles and tasks, hence situating health in a social, behavioral context. According to this definition, if the 'state of optimal capacity' is disturbed, health deviations occur. Parsons distinguishes four types of health disturbances:
1. Bad health in anatomic/pathological terms, as deviations in functioning of an organ or organ system. This is often seen as the traditional biomedical standpoint that expresses level of health-disturbance in type and size of diversions of the bio-physiological standards, mortality figures or morbidity figures (how many deaths or how many from which diseases). This technical conception is referred to by Parsons as 'disease'.

2. Health deviations as feelings of pain, fatigue, fear or indisposition; a popular topic in daily conversation and probably the most frequent reason to visit a doctor. According to this view, health status can only be judged by the sick individual himself. 'Illness' is the term Parsons uses for this type of health-perception.
3. Deviant health as limitation of functional or behavioral possibilities. According to this view, the functional status should indicate the level of health. This behavioral conception is called 'sickness'. It fits the strategy of measuring health through assessing the concrete functional consequences of health deviations.
4. Bad health as being under treatment of a health-care-professional. Frequency or intensity of the utilization of health care services indicates the health status at individual level or at population level [4,5].

Pain or nausea ('illness') cannot be experienced or judged by others than the individual in question. 'Diseases' are 'hidden processes which can only be understood as their observable signs are related to a body of knowledge about the way the human organism works' [6]. In other words 'diseases' are primarily judged by experts, in this case by health professionals. Behavior, however, can be observed and interpreted by the individual in question and by his surroundings, whether lay or health care professional. According to the 'sickness' conception, a normal level and type of functioning corresponds with good health, a decline in functional possibilities is seen as indicator of a decline in health. This view of health corresponds with the way people view their own health status and the health of others. It enables people to communicate about health in terms that are accessible to others. Professional health care is usually directed at 'disease': definition of the problem and the goal of treatment is stated in biomedical terms. The implicit assumption, however, is that treatment of underlying pathology will improve daily behavior ('sickness') or experienced health ('illness'). However, there is only a loose relation between a given state of health in 'disease' or 'illness' terms, and the level or type of associated behavioral changes ('sickness'). Moreover, an individual with a certain chronic disease might change his feeling state or functional status without changes in his 'disease' state of health. For instance, someone who lost a leg is in a chronic and stable 'unhealthy' situation. He might, however, adapt to his situation and live independently. On the other hand, someone with one leg might not adapt to the situation and be dependent on others for the rest of his life.

Illness, sickness and disease theoretically are three distinct concepts. In generating functional consequences of health deviations, however, they all three contribute. When an individual first contacts a health professional, his complaint will be stated in 'illness' or 'sickness' terms. And when a patient resumes a normal (or perceived optimal) level of functioning, treatment often is considered successful. Hence, the type and extent of behavioral changes is seen as an indicator for the level of underlying pathology (disease).
The relation between 'disease' and 'sickness', however, is not straightforward. The type of behavior or functioning can be distinguished according to their level of complexity or integration: the level of singular or isolated functions (e.g. leg or

hand movement), the level of complex integrated functions or skills (e.g. walking stairs, writing), the level of role performance (e.g. spouse, parent, or employee), or the most complex and abstract level of independent living as a 'human being'. The effect of a change at 'disease' level on functional status can be found at these four different levels of functioning. At every one of these levels a different kind and extent of changes or limitations will be associated to a given change in biomedical health status. At an integrated functional level, several singular functions cooperate, thus offering possibilities for compensation when one of these cooperating functions is affected. Hence, functional consequences, although large at the level of isolated functions, might be less than expected at a more integrated level. On the other hand, at a more complex and integrated functional level, the chain of effects from 'hidden process' to 'sickness' will be longer, less direct, and potentially subject to the influence of factors outside the directly health-related. Thus, a local functional limitation, resulting from a known 'hidden process', might cause functional consequences in unexpected areas of functioning. For instance, losing a leg has a massive and more or less predictable influence on an individual's possibility to move his leg or walk. However, the influence on a person's performance as chauffeur, employee, parent or spouse, is very hard to predict, and subject to many influences outside the area of health.
In figure 1.2 model the central 'negotiation box' depicts the confrontation of these different health definitions, the exchange of views in negotiations about the 'true' or valid definition of the situation. In paragraph 1.3.3 this process of negotiating will be elaborated in more detail.

1.3.3. Negotiation on valid health-related behavioral changes

When a person feels ill and expects not to be able to behave as usual, it is not enough for him simply to announce that he is ill; his evaluation of the situation has to be accepted by his co-actors. In other words, he has to receive 'provisional validation by the lay referral group' [6]. The definition of the true level of disability of the disabled person might not accord with the level of disability expected by 'the other'. If the disabled person and 'the other' agree on the expected level of disability, a situation of role-synchrony exists. The confrontation of visions and the exchange of views may lead to a revised definition of the health situation. This in turn results in a socially acceptable set of expectations that all participants agree upon. If, on the other hand, the disabled person and 'the other' differ in the expected level of health-related functional limitations and negotiations do not lead to agreement, their mutual expectations and behavior will differ, causing role-strain. The expectations connected with health deviations are formulated in Parsons' 'sick-role' concept. He developed this set of expectations associated with bad health to explain the behavior of people in a deviant health status and the reactions of their social surroundings. Parsons' health definition, as mentioned before, situates health in a social, behavioral context. If the 'state of optimal capacity' is affected, an individual might not be able to perform his usual roles in the usual way, and be forced to take on the sick-role. The expectations associated with the sick-role are:

(1) exemption from normal social role responsibilities, and (2) the individual himself is not to blame for his condition, and therefore must be taken care of. The third element is that health deviations are experienced as undesirable, and the individual wants to 'get well' (3). From this follows the fourth element: the obligation to seek competent help (4). Usually this help will be sought from a physician, and the person in bad health is expected to co-operate with this physician in the process of trying to get well.
From the description so far it might be concluded that the sick-role by definition is a temporary role. Health deviation is perceived as a temporary situation that will disappear when treated correctly by the ill person in cooperation with competent professionals. After recovery he has to resume his usual role-position and corresponding responsibilities and obligations. This poses a problem for chronically ill. The obligation to want to get well cannot be imposed upon them. By definition, their health condition will not resume the normal level after competent treatment, and consequently they will not be able to resume their normal roles. Hence, Mechanic [7] stated that the sick-role, as Parsons describes it, does not suit chronic conditions. The fact that the chronically ill and disabled lack the obligation to get well, means, in his view, that they play a different sick-role than the one described by Parsons. In reaction to this critique, it is stated that in the case of a chronic condition the sick-role does apply, albeit not in absolute terms but in a more relative form. The issue is not one of accomplishment of the goal of recovery, but rather one of approximation of this goal. Chronically ill or disabled people are not temporarily being exempted from more or less their total range of duties, but they are permanently exempted from a partial range of their duties [8]. Hence, it is not a question of presence or absence of the sick-role, but rather a question of a relative level which a person takes on, is forced to take on, or is allowed to take on the sick-role.
Following fig. 1.2, the level of sickness for an individual in a chronic deviant health situation results from the negotiations between the sick person and his professional and lay surroundings. An expected level of functioning, agreed upon by all parties, results from the negotiations as the 'true' level. Once agreement is reached a stable pattern of expectations will emerge.
At the onset of the disease, however, or in a situation where the health status changes, this pattern may have to be adjusted to a new health situation. Hence, once again, the process depicted in figure 1.2 appears to be circular. Continuously participants in the negotiations monitor the health status and check whether health changes (whether in disease, illness or sickness terms) occur that necessitate new negotiations on the validity of the present agreed level of functional limitations.

1.3.4. The outcome of negotiations

Thomas [9] developed a framework to classify the outcome of the negotiations about the level of health-related behavioral change, according to their validity. This model makes an inventory of the possible matches and mismatches between behavioral options of the disabled person and the 'others'. Friction or a problem is expected in all situations in which a misfit occurs between the chosen roles. To resolve the

strain, the disabled person or the 'other' will have to work towards a situation of role synchrony. When a person changes into a disabled person, the true normalcy option becomes impossible or at least less common. According to Strauss et al. [10], 'the chief business of chronically ill persons is not just to stay alive or keep their symptoms under control, but to live as normal as possible' despite the symptoms and the disease. Sick persons use tactics to make or keep a symptom invisible or, if it is visible, to reduce it to a minor status. They work hard at acting as normally as possible. Especially when the symptoms or the reactions of others are very intrusive, the sick person will experience serious role strain and 'the tactics for keeping things normal' have to be 'especially ingenious and elaborate' [10]. People who are ill may attempt to give a false definition when the truth might lose them jobs, status or 'face'. If a person can or will no longer deny the deviant health situation, the successful accommodation to lower levels of normality depends on the willingness and creativity of all engaged in the negotiations about the true or valid level of behavioral change. Negotiation will have to convince 'the other' of the real 'true' level of disability and accompanying level of behavioral change. On the other hand, non-sick persons, especially strangers, tend to overgeneralize the sick person's visible symptoms. They come to dominate the interaction unless the latter uses tactics to normalize the situation [10]. The optimal situation is that of valid role synchrony: a situation where the disabled person and the others hold the same definition of the level of disability or the level to which the health status impels to behavioral changes. In order to achieve this desired level of role-functioning, sick people will have to come to terms with their health status, and people who are significant to them (personally or professionally) will have to do the same. In these negotiations arguments might as well relate to aspects of illness as to disease or sickness.

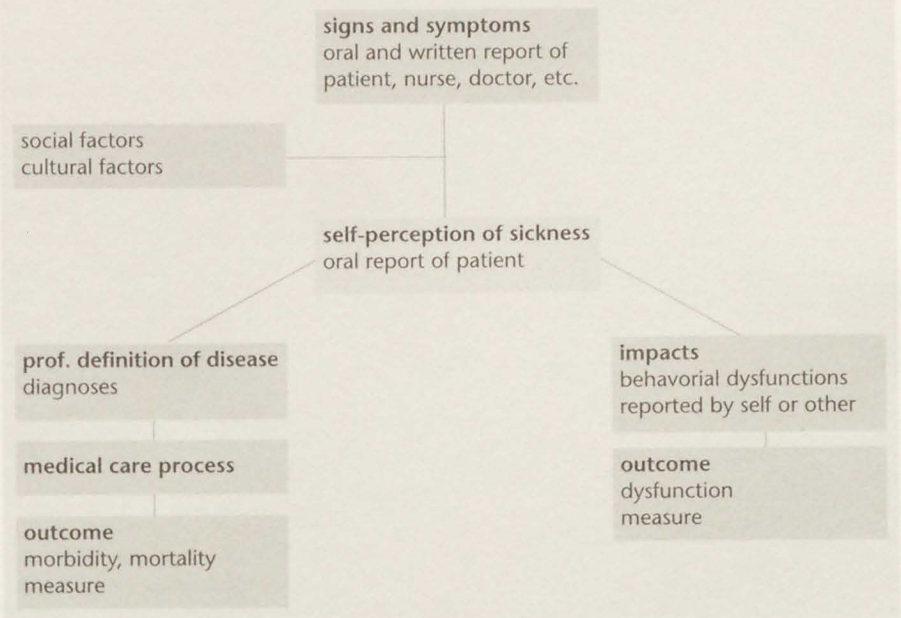
From the above theory and the model in figure 1.2 it can be derived that the actual health-related changes in behavior result from the confrontation of the evaluation of the situation by the disabled person and by 'the other', both based on their perception of the 'true' consequences of the health deviations or the 'true' level of functional status.
Our model depicts the process of determining whether or not, and to what level, to assume the sick role. The resulting behavioral changes represent the level to which the individual is allowed or forced to take on the sick role. During the period from falling ill to the recovery or optimal rehabilitation, a continuous process of negotiation is going on. All actors monitor the health situation and when one of them notices a change, new negotiations start on the level of health. Consequently, the acceptable level of behavioral deviations, or sick-role playing, is adjusted. Finally, when all parties agree that the patient has reached an optimal and stable situation, the acceptable health-related behavioral deviations (if any) are known and treatment ends. A stable situation will result, in which a new 'normal' situation exists. The former patient will then have to resume his life. If he is fully recovered and all actors agree on that, he will resume his usual roles, tasks and obligations. If a permanent disabling situation persists, his roles, tasks and obligations might be modified.

In conclusion, the model in fig. 1.2 emphasizes the social character of the consequences of health deviations. The expected level of behavioral status cannot be accurately predicted solely on the basis of a description of an individual's state of health, whether this is stated in anatomic/pathological ('disease') terms, in terms of feelings ('illness'), or in terms of medical consumption. Individuals with 'objectively' the same state of health will hardly ever show identical health-related behavioral changes, due to social factors that influence the relation between health and behavior. As the SIP is a measure of functional status, and it is used as an indicator of general health, this conclusion has implications for the application possibilities of the SIP and the interpretation of data provided by the instrument. In the next subsection, the Sickness Impact Profile is evaluated in relation to the model of health-related behavioral changes developed above.

1.3.5. The relation between the model and the SIP

The constructors of the SIP chose a behavioral definition of health. Hence, health is measured by assessing the influence of the health status on daily behavior or functioning. The model developed by Bergner et al. ([11], figure 1.3) in the construction of the SIP starts with signs and symptoms, recognized by an individual.

Figure 1.3. The SIP-Model (Bergner 1976)



This individual interprets these phenomena, depending on his social and cultural background, as consequences of a deviant health status. This has two consequences: on the one hand, professional caregivers are consulted. These caregivers start a care process aimed at reducing (the consequences of) health deviations. On the other hand, the sick person will still experience a certain level of behavioral dysfunction that he attributes to his sub-optimal health. The SIP, being a measure of health-related functional status, expresses this sub-optimal level of functioning.

When the SIP is viewed in relation to the general model of health-related behavioral change in fig. 1.2 (subsection 1.3.1), the concept the SIP aims to assess is situated at the right: 'health-related behavioral change'. One of the reasons the constructors of the SIP choose the behavioral or functional operationalization of health was that behavior can objectively be observed and hence this would supply an objective indicator of health. From fig. 1.2, however, it can be derived that social factors play an important part in determining the level and type of behavioral consequences of a given health status. Hence, a SIP score will not be a direct derivative of the 'actual' health status, but rather the result of a process of interpretation and negotiation in which the evaluation of the health situation from three points of view is the starting point. As the SIP, and related instruments, are often used in health care evaluation studies, this social factor should be taken into account when score changes or score differences are interpreted. Changes in reported health-related functional status (= SIP scores) might not result from changes in health status in the narrow ('disease') sense, but from changes in attitudes, interpretations or perceptions of one or more of the persons participating in the negotiations. In other words, 'disease' is not necessarily connected with behavioral changes; and behavioral changes, although perceived as health-related, are not necessarily and directly determined by a disease. Although the patient himself fills out the SIP, in the process of negotiating about the true level of behavioral changes, professional caregivers and 'lay others' influence his perception of what behavioral changes can rightfully be ascribed to his health deviation.
An example of this is a person suffering from cancer, with a tumor that can be effectively removed and that does not cause any subjectively noticeable signs or symptoms. If the tumor is removed, the 'disease' status shows dramatic improvement, while no health-related behavioral changes occurred. However, the fact that this same person is told that he has had cancer might have a drastic influence on his behavior or on the attitude of his co-actors towards him, although in a biomedical sense, these behavioral changes are not related to the tumor.

1.3.6. Implications of the health behavior model for the SIP

The model developed above depicts a process of negotiation between the person in a deviant health status, his social surroundings and professional caregivers about what behavioral changes can rightfully be attributed to the health deviations. These negotiations about the true unavoidable behavioral changes, partly explains the broad variability of behavioral changes connected with a given health problem.

Moreover, the fact that negotiations have to take place before the true level of health-related functional impairment is reached, has implications for the time-schedule of a study in which the functional consequences of health problems or health care is evaluated. The final effect of treatment on functional status, for instance, will not be apparent immediately after the treatment, but only after participants in the negotiations have had the time to reach an agreement on the acceptable new level of functional consequences.

Another implication of functional status as a result of negotiations is that a change is not to be expected merely as a result of the treatment of the bio-medical health status. Unless drastic and obvious changes in health take place, no changes in the agreed functional status are to be expected. Hence, a treatment at disease level might not lead to an enhanced functional status, although the 'hidden process' was successfully suppressed. An unaffected SIP score might point at an ineffective treatment; however, it also might be caused by a recalcitrant and inert behavioral pattern, based on a strong belief in an inadequate definition of the situation that is confirmed by the patient and his co-actors. An actual change in SIP score, on the other hand, might be caused by treatment and resulting changes in the disease situation. It might also be caused, however, by a change in the interpretation of the situation by the sick person or other participants in the negotiations. This last phenomenon can be seen in the chronically ill that adapt to their situation, and find ways to behave as they want in spite of their sub-optimal health status.

The model also has implications for treatment strategies in populations with chronic diseases. In a situation where health care is primarily directed at minimalizing functional limitations and not at healing the ailment, as in chronic disease, substantial attention should be directed at the social aspect of functional status. For instance, by explicitly involving the spouse of a patient in treatment to stimulate an optimal attitude towards the patient. This might enhance and consolidate the treatment results. It also follows from the model that health care professionals should be aware of the influence of their own subjective and objective evaluation of, and attitude towards the patient's situation. The verbal and non-verbal information they supply is an important element in the negotiations between patient, lay others and professional care providers, and hence might be brought into action to influence the functional status.

Naturally this possible influence of social factors is restricted by the absolute physical restrictions: someone without legs will never run up and down stairs. Further study could be done, though, into the empirical tenability of the model. This could be done by administering the SIP to a patient with a chronic disease and his professional and lay surroundings, several times throughout the trajectory of a disease. Based on the model it could be hypothesized that in the acute phase of a disease SIP scores might differ greatly because it is not yet clear what exactly the matter is, and all involved have different expectations and interpretations of the situation. In the period of treatment, negotiations about and adjustments of the expected true level of health-related functional status will start. When the disease has reached a stable phase, and the patient and his social surroundings have had time to adjust, experiment and negotiate, a higher level of agreement between SIP scores will be

attained. Every time a shift in the disease situation is noticed by one of the participants, new discrepancies in SIP scores might occur, and thus new negotiations will start. When the change is considered large enough, this might, in time, lead to agreement on a new level of health-related functional status, and consequently a new SIP score. As several parties have to be convinced of the necessity of adaption, it is to be expected that only relatively large and distinct changes in the disease situation or in the perception of this situation will lead to adjustment of the generally accepted 'true' level of health-related functional status. It can also be expected therefore that the SIP score of someone with a chronic disease after the initial, acute phase, has a stable character. Hence, changes in SIP scores are only to be expected when large and obvious changes in the perceived health status occur.

The processes described above concern interpretation and negotiation relating to the expected behavioral consequences of health deviations. The description of these processes in relation to the use of functional status measures in general and the SIP in particular, however, can only be grounded on theory. Empirical studies into these processes in relation to health measurement are not found. Within the framework of the SIP project the model of health consequences provides a theoretical frame of reference for the interpretation of SIP findings. The model itself will not be explicitly tested for its empirical tenability. At this point, however, it can be concluded that the interpretation of health-related behavioral status as an indicator of health is less straightforward than it is usually thought to be. The interpretation of this type of health indicator in combination with the model developed above, potentially supplies very useful data for research in health care. Apart from conceptual clarity, however, the psychometric properties of a measurement determine whether or not it is a useful tool in research. In the next section, therefore, the psychometric characteristics of the SIP are presented as far as they could be derived from the literature.

1.4. Psychometric characteristics of the SIP

In order to describe the state of the art of the SIP, an extensive review of international methodological literature was performed. It appeared that the SIP is used in a number of different languages. Translations were found into British English, Swedish, German, French, Danish, Norwegian and Dutch [3]. Psychometric information, however, was only found on the original American SIP, on the English, on the Swedish and on the Dutch version. This section presents the information found.

1.4.1. Reliability

Reliability is concerned with the level to which the information supplied by an instrument is influenced by random error. Usually two main types of reliability are distinguished: test-retest reliability, and internal consistency. Test-retest reliability concerns the level to which an instrument supplies identical information when it is

used in an identical situation. The test-retest reliability of the SIP appeared to be good to very good for all translations. Both the SIP total score and the two dimension scores could be accurately reproduced in a test-retest procedure. Correlation coefficients between two successive administrations ranged from 0.75 to 0.92 for the total score, and from 0.79 to 0.91 for the two dimensions. These figures imply that when the SIP is used twice in an identical situation, the information obtained is also identical to a very high degree. Hence relatively little random error or random variation is incorporated in a SIP score. The second type of reliability usually reported is the level of internal consistency. An instrument is internally consistent when all items in the instrument have bearing on the same concept; in other words, when all items measure an aspect of the total concept measured. Insufficient internal consistency implies that the instrument does not measure one coherent concept, but that several different subconcepts are measured [12]. The internal consistency is usually expressed by means of Cronbach's α [13]. This indicator theoretically ranges from 0 to 1.0. An α of 1 indicates that all items measure exactly the same concept. This would mean that every item supplies the same information as all the others. In this case, hence, more than one item would not be needed, and an instrument of 136 items might be considered largely redundant. An α of 0.80 is usually accepted as a desirable or sufficient level of internal consistency [14]. For the SIP, throughout the literature, good to very good levels of Cronbach's α for the total list are found (range 0.91 to 0.95). α's for the dimensions and categories also indicate sufficient internal consistency (range 0.84 to 0.90 and 0.45 to 0.90 for the dimensions and categories respectively) [3]. In sum, it can be concluded that based on findings concerning test-retest reliability and internal consistency, a SIP score contains relatively little random error; hence the SIP is a reliable instrument.

1.4.2. Validity

When the validity of an instrument is judged, it is assessed to what degree the information supplied is biased by systematic measurement error. In other words, does the instrument really measure what it is supposed to measure, or does it systematically aim at another (possibly related) concept [12]? In methodological literature, a number of diferent types of validity are found. Three basic types of validity can be distinguished that are relevant in this situation: content validity, criterion validity and construct validity. Each of these types has bearing on another aspect of validity, or on another technique of looking at validity.

Content validity is concerned with the extent to which the items of the instrument adequately represent the total property being measured. Inspection of the items by specialists and judging the item-selection procedure might indicate the level of content validity of an instrument. The SIP is meant to be a measure of general health: hence it should incorporate all three aspects of health mentioned in the WHO definition: physical, mental and social health. During development of the instrument, the items were gathered in a scrupulous and extensively described way to ensure a broad scope on possible consequences of health deviations. In general,

correlations among categories appear to be relatively low, indicating minimal overlap [15]. No major aspects of health-related functional status are mentioned as missing in the literature used in this study. Hence, it was concluded that the SIP has a sufficient level of content validity [3].
Criterion validity is assessed by comparing findings of the instrument under study with findings of instruments known or believed to measure the same concept [16]. As the SIP is a general health measure, the relation with other measures of general health are studied. Correlations of SIP scores with self assessments and doctor assessments of health are found to be sufficient in order to assume clinical relevance and criterion validity of SIP assessments. Also the relation between SIP scores and assessments by other general health questionnaires was found supportive for the criterion validity of the instrument [3].
The construct validity of an instrument is judged by the level to which the findings of the instrument under study fit the theoretically derived network of relations of the instrument with measures of other (relevant) concepts. To be able to study this type of validity, the theoretical position and content of the concept being measured and the relations with the network of related concepts within which it is situated should be explicitly described [12]. During the development of the SIP, however, theories on the concept 'health-related functional status' were not made explicit. The theoretical considerations described in section 1.3 is the first attempt known to the authors, to build a theoretical frame of reference for the SIP. Hence, construct validity in the strict sense has not been studied. When information has been presented on the construct validity, it usually is based on individual studies in which the SIP is *used together with one or more, more or less related measures.* When the relation between the SIP and these other instruments is in accordance with the expectations, this is usually seen as supportive for the construct validity. Based on this kind of ad hoc evidence it can be concluded that the construct validity of the SIP is supported: relations between the SIP total score and scores on both SIP dimensions on the one hand and different kind of related measures on the other, as far as can be derived from published studies, are usually congruent with theoretical expectations. The SIP psychosocial dimension shows stronger relations with measures of psychosocial aspects of health, and the SIP physical dimension is usually more closely related to physically oriented instruments. The SIP appears capable of describing and delineating groups of illnesses within varying test populations [3].

A different way to evaluate the construct validity is to study the internal structure of the concept measured by the instrument. The fact that twelve categories are distinguished suggests that twelve different aspects are to be found within the construct 'health-related functional status'. However, no information was found on this topic. In other words it is not clear whether the internal (categorical) structure of the SIP is the valid structure of health-related functional status. As the concept measured by the SIP is an intersubjective operationalization of health, scores obtained *from the individual under study and scores obtained from persons familiar with the* functional status of the individual under study (proxy respondents) should be similar. Hence, the relation between SIP scores obtained from individual experiencing

health deviations and SIP scores of proxies of these individuals judging them by means of the SIP might be used as an indicator of construct validity of the SIP. Although some researchers have investigated the possibility of administering the SIP to proxy respondents, validity aspects of proxy responses have not yet been adequately tested. This topic of proxy SIPs might be part of a project directed at the theoretical tenability of the behavioral health consequences model developed above. As stated before, however, such a study aimed at the corroboration of the model in figure 1.2, is outside the scope of the SIP project.
The preliminary conclusion as far as construct validity of the SIP is concerned is that indications of construct validity are found, and no firm evidence is found against construct validity. Further study of this topic, however, is certainly needed before a final conclusion can be formulated. In this further study, the model developed in section 1.3 might serve as a basis for the formulation of theoretical hypotheses.

1.5. The development of a short SIP-version: the SIP68

Using the SIP in several projects, the impression arose that the instrument is rather long. In the literature review, it was found that the length of the instrument is also considered by other researchers one of its major drawbacks [3]. Compared to related instruments the SIP indeed is very long and, therefore, places a relative large burden on respondents that have to complete the list. Moreover, for research projects that 'just' need an assessment of the health related functional status and no detailed individual profile of functional status, the SIP is rather inefficient. At the same time it was found that the validity of the twelve category structure had never been explicitly studied. Hence, the relevance of all twelve categories for measuring the health-related functional status had not yet been empirically studied. Therefore, it was decided to study the internal structure of the instrument. Insight into the major components that would be distinguished within the total concept of health-related functional status would increase the construct validity of the SIP. This internal structure in turn could subsequently serve as a basis to distinguish between relevant items and less relevant items. This distinction would be the basis to select the most relevant items that would be selected for a short SIP version. As mentioned above, a large SIP database had already been gathered. This database was used to develop a short version of the instrument. The aim was to develop an instrument with the same generic application possibilities and the same level of psychometric properties, but containing substantially fewer than the original 136 items. As a starting point for this exercise, the twelve official SIP categories were used.

1.5.1. Item weights

The weights used in calculating the SIP136 score are weights expressing the differences in impact between the items in the original twelve categories. To assess the possible loss or bias of information caused by selecting items without accounting for the weights, the amount of influence of the weighing procedure on the score was

studied. For each category for the two dimensions and for the total instrument, scores were calculated in two different ways: the traditional weighting procedure, and a procedure in which every item has a weight of 1. In the second procedure the number of items checked in a category, a dimension or the total list is the score obtained. The correlation coefficient between weighted and unweighted scores was very high (range: 0.94 to 0.99, median 0.98) [17]. From this it was concluded that no essential bias would result from using a scoring procedure without weights.

1.5.2. Data and methods used to select items

The data used in this phase of the project were gathered in several studies in which the Dutch translation of the SIP was administered. Ten different diagnostic groups are represented in the total data base, containing over 2500 respondents. To prevent bias by unbalanced representation of diagnoses in the total population, samples were taken from every diagnostic group. This resulted in a group of 835 respondents with a balanced representation of 10 different, mainly chronic or lengthy health problems [18]. This population was used in further analyses.

The first step in reducing the length of the SIP was dropping the category 'work'. This category does not apply to most of the respondents because they did not work outside the house before their health problem occurred (housewives, retired persons, students). Moreover, due to differences between social security systems in the Netherlands and the U.S.A., this category did not supply relevant information in the Dutch situation. The item selection procedure was continued by judging the relative relevance of the items in two ways: It was hypothesized that if the concept of health-related functional status is adequately operationalized by the categories of the SIP, it would be possible to find evidence for the validity of this categorical structure in empirical data. Principal components analyses (PCA) [19] was applied to discover the structure of the data set. This technique reveals the main independent dimensions of variation in group characteristics. It provides components ('factors') that describe relatively independent subgroups of variables within the data set. In doing so, sub-concepts within the concept measured are revealed, enhancing conceptual clarity and thus the construct validity of the instrument.
The relevance of items in the factor structure is indicated by means of factor loadings. By selecting the highest loading items in every factor, the most relevant items from the factorial structure will be obtained. A factor loading of .40 indicates that 16% of the variation of that item is involved in that factor. Items with a factor loading of less than .40 on any factor are considered less relevant for the instrument. The other criterion to judge relevance of items was the skewedness of the response pattern. As the aim of the instrument is to distinguish groups within a population, extremely skewed items, from a research point of view, can be considered less relevant. Items applying to 10% or less, or to 90% or more of every diagnostic sub-population, were considered skewed, and were removed from the list.
The first PCA did not render evidence for the empirical validity of the a priori twelve-factor structure (eleven after deleting 'work'). After removing the skewed items, an

interpretable six-factor structure was found. After deleting low loading items, 68 items remained, divided over 6 factors. The structure of the selection was tested for its robustness by comparing factor solutions in different sub-populations. Cattell's salient similarity index indicated that the same structure was present in all sub-populations. As the selection contains 68 items from the SIP, the short instrument will henceforth be called SIP68. For reasons of clarity, 'SIP136' will be used when the original instrument is meant.

1.5.3. Result: the SIP68

The procedure described in 1.5.2 led to a selection of 68 SIP items, divided over 6 categories (see fig 1.4). As these categories were based on findings of a PCA it was assumed that these six categories represent sub-concepts of 'health-related behavioral status'.
The first category was interpreted to represent the aspect of 'Somatic Autonomy' (SA). This category assesses the level to which the respondent is autonomous in his or her basic somatic functions like getting dressed, standing, walking, eating. A higher score on this category means a high level of dependency in this area. As a heading for the second factor 'Mobility Control' (MC) was chosen.

Figure 1.4. SIP68 categories and item examples

Somatic Autonomy
- I get dressed only with someone's help.

Mobility Control
- I walk shorter distances or stop to rest often.

Psychic Autonomy & Communication
- I have difficulty doing activities involving concentration and thinking.

Social Activity
- I am cutting down the length of visits with friends.

Emotional Stability
- I act disagreeable to family members, for example, I act spiteful, I am stubborn.

Mobility Range
- I am not doing any of the shopping that I would usually do.

This category describes behavior that indicates the level to which the respondent has control over his body. Half of the items in this category are related to walking, the others have to do with hand- and arm-control. A high score on MC indicates a relative low level of control over body movements. 'Psychological Autonomy and Communication' (PAC) is the third SIP68 category. Here behavior is described that is associated with the level to which an individual is able to function without help of others in areas of mental functioning, including the possible health deviation impacts on a person's (verbal) communication. The fourth category was interpreted as an indicator of the consequences of a health deviation for 'Social Behavior' (SB). Functioning in relation to other persons (spouse, children, others in general) is described in statements on, among others, sexual activity, visiting friends and activities in groups of people. Two items concerning eating and drinking also are found in this category. Apparently these activities are more directly related to their social context than to other (eg. physical) aspects of health. Someone obtaining a high score on this category will experience functional limitations in social functioning. The fifth category is headed 'Emotional Stability' (ES). This factor estimates the effect health status has on the emotional status of the respondent. Among others, irritability and acting disagreeably with oneself or others are mentioned as possible functional impacts of bad health. A high score on ES points to a lack of stability in, or control over the emotions. The final factor, 'Mobility Range', is concerned with the influence of health deviation on a number of usual tasks like shopping, housecleaning and taking care of personal business affairs. It is not the level of control over motor functions that is described, but the range of actions to which the respondent has (limited) disposition, given the level of motor control.

All these six aspects of functional status, motor or mental possibilities might be affected by health deviations. This might result in changes of behavior. The extent of these behavioral changes (level of score on the SIP68) is an indicator of the severity of the health deviation. As described in section 1.3, the perception or the interpretation by the person in question and his surroundings of the level and/or type of health deviation also influences the resulting functional consequences. It might well be, for instance, that the spouse of someone who had a heart attack urges this person to rest and to restrict physical activity because 'it might happen again'. When this results in the decision to limit household activities or social activities like visiting friends, this would lead to an increased score on categories Social Behavior and Mobility Range. A cardiac rehabilitation program on the other hand might stimulate the person in question to explore his 'true' functional possibilities, which could result in an increase of social and motor activities. In spite of the fact that the objective cardiac condition does not change, this hence might lead to a decrease in score on the categories Social Behavior and Mobility Range. The same mechanism might be found in somatic functioning (Somatic Autonomy), motor functioning (Mobility Control) and mental functioning (Psychic Autonomy and Communication). For deviations in emotional functioning it might be that a perfectly normal and adequate emotional reaction of 'a patient' is interpreted by

his surroundings as connected to or a result of the deviant health status. If this perception is adopted by the patient, this would result in an increased score on Emotional Stability. The opposite also might occur: due to a health deviation a patient is more emotional than before, but he does not interpret this as a result of the health deviation. In this last case therefore the score on Emotional Stability falsely will not increase. Just as in the interpretation of a score on the SIP136, social as well as physical determinants should be taken into consideration in the interpretation of the score on the SIP68.

1.5.4. Psychometric characteristics of the SIP68

To obtain a preliminary impression of the reliability and validity of the SIP68, these characteristics were explored using the instrument as it is derived from the available data obtained using the SIP136. As respondents have had the possibility to chose from 136 instead of 68 items, extracting the SIP68 from the administration of the SIP136 might influence findings on the reliability and validity of the SIP68. However, these findings still supply an indication of these characteristics of the SIP68 as if it had been administered as a separate instrument. The internal consistency of the SIP68 was judged by means of Cronbach's α [13]. For every category and for the total list the Cronbach's α was above the level generally accepted as sufficient of 0.70 (0.92 for the total SIP68), indicating that all items have relevance for the assessment health-related functional status.
An indication of the validity of the SIP68 can be found when the relation between SIP68 scores and scores obtained on the SIP136 using the original items weighting procedure is studied. The SIP136 is then used as the criterion against which the SIP68 is judged (criterion validity). To study this relation, four regression formulas were calculated. Regression analysis allows one to study whether it is possible to predict the score on one variable using information from one or a number of other variables. The level to which the prediction is correct is expressed in the 'fit' of the regression equation. A good 'fit' indicates a close relation between the information provided by the variables used in the analysis. First, the SIP136 total score was predicted from the SIP68 total score. Next the score on the SIP136 was predicted using the scores on the six SIP68 categories separately. Finally the original physical and psychosocial dimension scores were predicted from the six SIP68 category scores. From the R^2's it can be concluded that the 'fit' for all four regression equations was good to very good (0.94, 0.96, 0.96, 0.88 respectively). This indicates a very close relation between the information provided by both SIP versions. Apparently only a very small amount of information from the SIP136 total and dimension scores is lost in the selection procedure. The same regression analyses were performed using that part of the population that was not used to develop the SIP68. Findings in this group were almost identical to findings in the group used to select the items. To test the robustness of the factor structure, principal component analyses (PCAs) were performed on SIP data from respondents that had not been used to develop the SIP68. This population was divided into four different diagnostic subgroups. Within each of these subgroups PCAs were performed and the resulting factor

solutions were compared with the SIP68 categories. Cattell's salient similarity index s [20], an indicator of similarity between factor solutions, indicated that there was no significant difference between the solutions found in the different diagnostic groups and the category structure of the SIP68. This finding supports the validity of the SIP68 category structure.
To compare the internal structure of the SIP136 and the SIP68 at the level of dimensions, a second order PCA was performed on the category scores of both instruments. These analyses revealed that both instruments contain two dimensions that appear to cover similar aspects of functional status: one dimension concerned with physical aspects of functioning, and a psychosocial dimension. Both the SIP136 and the SIP68, therefore, can be seen as indicators of a broad health concept.

As stated above, hardly any information contained in the SIP136 appears to be lost in the selection procedure. This leads to the question what information is to be found in the items that were not selected. Items from the category 'work' are not considered relevant for our population and for the Dutch situation, and hence are removed on appropriate grounds. Items with a very skewed answering pattern do not differentiate within the population and, therefore, these items also are rightfully removed. The 40 items removed on the basis of their low factor loading, however, might contain useful information on the health-related functional status. However, it was found that hardly any information was lost in the selection process. Hence, it is to be expected that the 40 remaining items contain information that also is supplied by other SIP136 items. The 40 items, thus, could be redundant or double items from the SIP136. To study the content of this list of 40 items, a principal components analysis (PCA) was performed. Only one factor appeared to be interpretable as a possible representation of a theoretically distinguishable aspect of health that is not represented in the SIP68. This is a factor containing 5 items representing social isolation and/or loneliness. However, adding these five items to the SIP68 only very marginally changed the SIP68 scores. When subsequently a PCA is performed on the SIP68 including the social isolation items, the five social isolation items are divided over factors Social Behavior and Emotional Stability. Apparently the aspect of social isolation is not as a separate factor present in empirical data. Probably this concept is of a different theoretical level of health or health-related behavioral status and it is a result of behavior described in Social Behavior and Emotional Stability. With respect to the accuracy of the SIP68 selection, however, the analyses of the items that were dropped support the validity of the SIP68 as an alternative to the SIP136.
At this point it was concluded that based on the findings in this study, the SIP68 is a very promising short generic alternative for the original Sickness Impact Profile. A crucial point in the validity of findings during the development of the SIP68 was that the data found on the SIP68 all are based on information gathered by administering the long SIP136. This implies that the context of the SIP136 might have influenced the findings on the SIP68. To investigate the psychometric characteristics of the SIP68 when it is administered on its own, two studies were performed. Findings in these studies will be described in the next section.

1.5.5. SIP136 and SIP68, complete coverage versus parsimony.

The SIP136 is renown for its broad scope and complete coverage of the concept of health related functional status. Almost by definition, reduction of the number of items in the list will lead to reduction of this coverage. Hence, the question is raised *on what grounds it is to be justified to use a short version of the SIP.*

First, the possible applications of the instrument are considered. When the SIP136 was developed it was intended to be used in 'measuring the outcome of care in health surveys, in program planning, in policy formation and in monitoring patient progress'. *These aims require data on functional status of different levels of detail.* Following an individual patient in his or her progress in a cinical situation, requires a detailed, checklist-like instrument in which many different details or examples of behaviour that might be affected by the health situation are explicitly checked. The fact that the SIP136 is meant to be a generic instrument leads to a very broad catalogue of possibly affected behavior. This catalogue is very well suited to evaluate the individual functional status in a clinical situation. On the other hand, in conducting a health survey or in formulating a general health policy, the researcher needs an instrument that provides a general assessment of the functional status of a population, or of the differences in functional status within a population. Usually, in such a situation, data of a more global or abstract level is sufficient or even more efficient. This difference in use is mirrored in the possible levels of scoring the SIP136. The lowest level is that of item scores. At individual level it might be interesting to know whether a specific behavior (still) is affected by the health status of tha individual. Hardly ever scores at this level are found in research literature. The next level is information expressed in category scores. These offer the possibility to provide a profile of functional status, consisting of the twelve different aspects (categories) that were defined by the constructors of the instrument. SIP-data at category level is found in literature, but hardly ever category scores are used to express the functional status. Most frequently the functional status measured by means of the SIP is expressed in dimension-scores or by means of the totalscore. Using dimension scores functional status is viewed as containing two aspects: physical functioning and psychosocial functioning. The physical dimension score is composed of categories Bodycare and Movement, Ambulation and Mobility. The psychosocial dimension score is calculated from categories Social Interaction, Alertness Behavior, Emotional Behavior and Communcation. The other five categories are not used to calculate dimension scores, apparently the model of functional status containin twelve aspects is not congruent with functional status build from two dimensions. Finally, the total SIP136 score is a representation of all category scores into one assessment of functional status.
As mentionned before, SIP136 scorelevels that are most frequently found in research literature are dimensionscores or total scores. The explanation for this fact might be that these levels of abstraction better fit most research questions asked. It might be that the constructors of the SIP136, by defining twelve different categories, developed a model of health related functional status that is to complex or to differentiated to fit reality as it is viewed by most researchers.

This consideration about the model of functional status underlying a functional status measure, leads to the question of parsimony of the instrument or its underlying model. How many items or groups of items are needed to give an adequate assessment of the health related functional status? On the one extreme the constructor of an instrument wants to check every possible aspect of the concept measured, resulting in an infinite list of items. On the other extreme the instrument exists of only one item asking directly for an assessment of the construct measured ('how are you?' or 'how is your health related functional status?'). The first option will lead to an inefficient instrument. If the number of items describing different details of behavior rises, more items will be irrelevant or redundant for an individual filling out the list. Moreover, within a large number of items aimed at the same construct, relatively many items will be conditionally related, and the score on one item will predict the scores on other items vice versa. Hence, also the level of redundancy will rise. When used in a population, the abundance of items and the individual variance will necessitate aggregation of items into more or less homogeneous subgroups to prevent that a hazy and non informative picture arises of the total construct. The second option on the other hand (only one item), will lead to a rather poor or 'thin' description of the construct. One might have obtained a very global assessment of the construct aimed at, but information on the content of that construct or its internal structure will not be available. Hence, the population measured might be devided into levels of functional status, but as no insight is provided on the aspects within this construct, this is rather 'flat' or uni-dimensional information.

The optimal level of parsimony to describe functional status lies somewhere between the endless list of items on the one hand and a one-item-instrument on the other. Where the equilibrium between these two extremes is found depends on the intended use of the instrument, and the level of abstraction suitable for the application. Absolute criteria to decide when this equilibrium is reached are not available, the specific research question and the subjective judgement of the researcher allways will play a part in the decision to choose an instrument or not. This choice implies a trade off between two aspects of validity, content validity and construct validity. When the accent is on content validity, complete coverage will have highest priority. When, on the other hand, construct validity has priority the efficient en consistent representation of the construct measured will be the primary goal. Some well known measurement instruments, hence exist is several lengths. The Barthel index is one of the best known, but also a short SIP version has been developed to measure health related functional status in populations suffering from back pain, the 24-item Roland scale.

As discussed above, the original SIP136 was intended for use at different levels of abstraction (evaluation of functional status both at individual and at population level). This ambivalence resulted in an instrument that, in terms of economy or parsimony, is not optimally suited for use at population level. In constructing the SIP68, the lowest level of abstraction (136 individual items describing behavior) was the starting point. As the aim was to develop an generic instrument to be used as indicator of functional status at population level, we sought after a description

of the total construct for every of our subpopulations in less than 136 scores and more than only one total score. As described elswere the twelve category model of the SIP136 did not fit our data. The six elements defined by the six categories of the SIP68, however, were present in the total group as well as in all different diagnostic subpopulations. This six factor model of functional status, hence, appeared to fit the aims of a generic functional status measure at population level. For this reason, this six category structure was the basis for item reduction.
In conclusion, it can be stated that, when a detailed description of the (progress of the) functional status of an individual patient is wanted, the SIP136 would be the instrument of choice because of its level of detail. When, however, the health related functional status is to be assessed at population level, or the functional status of different diagnosic groups is to be compared, the SIP68 would be prefferred above the SIP136.

1.6. Reliability and validity of the SIP68

To investigate the difference between psychometric characteristics of the SIP68 extracted from administration of the SIP136 and these characteristics of the separately administered SIP68, a project was started in which the reliability and validity of information provided by both ways of administering the SIP68 was compared [18]. A second project was performed in which only the SIP68 was administered, primarily directed at evaluating the psychometric characteristics of the SIP68 [21].

1.6.1. Data and methods

In the first project the population consisted of visitors to an outpatient rheumatology clinic at a general hospital. Fifty-one respondents filled out the SIP68 at the outpatient clinic (t1). The mean age of this group was 57 years, 69% were female, 47% suffered from rheumatoid arthritis and 53% from various other rheumatic diseases. Two days after the first assessment (t2) the respondents again filled out the SIP68 (at home) and returned it by mail. The third assessment moment (t3) was fourteen days after t2. At t3 the respondents were visited at home and filled out the SIP136 in the presence of a researcher. The last assessment took place two days after t3, and again the respondents filled out the SIP136 and sent it back by mail. At every assessment an additional questionnaire provided self-assessments of functioning (physical, psychological and social), happiness and the severity of the disease on five point Likert scales.
Rheumatic diseases usually are relatively stable (at least over a period of two weeks). Hence, no substantial changes in the health-related functional status are to be expected over a period of two weeks. Comparing scores on the SIP68 from the first two assessments with SIP68 scores derived from the administration of the SIP136, permits us to study whether the context of the SIP136 substantially influences the score on the SIP68. By studying the relation between SIP68 scores at all four assessments, the test-retest reliability, or reproducibility of the SIP68 can be assessed.

Because the administration took place in different situations (at the outpatient clinic with an interviewer present, at home without anybody present, at home in the presence of an interviewer) the influence the way of administration has on the SIP68 score can also be studied.
In the second project the population existed of 315 persons (75.4% male) with a spinal cord injury, who had a mean age of 39.4 years and who were living in the community at the time of the study. A complete quadriplegia existed in 21.7%, 20.4% had an incomplete quadriplegia, 29.2% was completely paraplegic and 28.6% incompletely paraplegic. Most respondents (60%) were wheelchair-dependent. The time after injury ranged from 1 to 7 years. Of all respondents the level of lesion and the vocational situation was known. Apart from the SIP68, two other instruments were administered: the Barthel Index (BI) a well-known measure of functional independence in activities of daily living, and the Life Satisfaction Scale (LSS) measuring the satisfaction with life in general and with eight specific domains (eg. self-care, family relations). The relation between the SIP68 and these other instruments supplies information on the validity of the SIP68 as a measure of health-related functional status in this population.

1.6.2. Reliability of the *SIP68*

The test-retest reliability of the SIP68 was assessed in the first project by calculating the intraclass correlation coefficient (ICC) [22]. This reliability coefficient was calculated over the scores obtained at all four assessments. The theoretical possible range of the ICC is from 0 to 1. As the functional status of the respondents was considered to be stable across all four assessment moments, an ICC close to 1 would support the reproducibility or the test-retest reliability of the SIP68. For the total instrument an ICC of 0.97 was found, indicating very good test-retest reliability for the total SIP68 score, over all four assessments. The ICCs found for the separate category scores within the SIP68 ranged from 0.90 to 0.97. This also can be considered as an indication of good to very good test-retest reliability or reproducibility. From these findings, it can also be derived that the information gathered by means of the SIP68 as a separate instrument is not essentially different from information gathered with the SIP68 as a part of the SIP136. Test-retest reliability also was evaluated at item level. The Jaccard's Similarity Ratio (JSR) [23] is an indicator of the level to which items checked at one occasion, also are checked on the second occasion. The JSR indicates considerable agreement between items checked at t1 and t2 and t3 and t4 respectively. Very little difference appeared to exist in the agreement between t1 and t2 on the one hand and t3 and t4 on the other. Respondents are apparently consistent in the items they check, and the reliability of the SIP68 is not influenced by the context of the SIP136. This finding offers the possibility to connect databases that contain SIP136 results with databases gathered in studies that used the more efficient SIP68, by selecting dichotomous answers to SIP68 items from the SIP136 databases and using these to calculate SIP68 scores. Cronbach's α was calculated to judge the internal consistency of the SIP68 in both

the rheumatoid project and the spinal cord injury population. For the total list in both studies a Cronbach's α of 0.90 or higher was found, indicating (very) good internal consistency. The α's at category level were somewhat lower (range 0.68 to 0.80) but still sufficient to assume an acceptable level of internal consistency.

1.6.3. Validity of the SIP68

The validity of the SIP68 was studied by comparison of findings of the SIP68 on the one hand and findings of the Barthel Index, the Life Satisfaction Scale, and data on level of lesion and vocational situation, on the other. The level of lesion of a spinal cord injury indicates the expected level of locomotory limitations. In general, the higher the lesion, the more locomotory limitations are to be expected. The SIP68 total score, as well as 5 out of the 6 category scores show a significant correlation with the level of lesion in the expected direction. Correlations for the category scores differ: they range from 0.31 for psychological aspects (category PAC) to 0.72 for category Somatic Autonomy. As level of lesion is expected to be more closely related to physical aspects of functional status than to psychosocial aspects, these findings support the construct validity of the SIP68. The correlation between the SIP68 total score and the level of lesion was 0.59, indicating an indirect but relatively strong relation. Less high correlations were expected and found between the SIP68 and life satisfaction (score on the LSS). The LSS total score showed a correlation of 0.52 with the SIP68 total score and correlations between LSS and SIP68 category scores range from 0.22 to 0.53. As the SIP68 aims at measuring a construct that is related but not very closely related to life satisfaction, these findings also support the construct validity of the instrument. The SIP68 and the Barthel Index (BI measure of functional independence in activities of daily living) [24] are supposed to be rather closely related. Correlation coefficients between these instruments indeed appeared to be relatively high: 0.74 for the total scores. A correlation coefficient of 0.91 was even found between SIP68 category Somatic Autonomy and the total score on the Barthel Index. Apparently Somatic Autonomy supplies information on an aspect of functional status that is very closely related to the concept measured by the BI: functional independence in activities of daily living. The SIP68 apparently incorporates this concept, but as total instrument aims at a broader conceptualization of functional status. These findings all support the validity of the SIP68 as a measure of health-related functional status. To verify the validity of the internal categorical structure of the SIP68, it was studied whether the six-category structure of the instrument could be found in data collected using it as a separate instrument. A PCA on the data from the rheumatoid population was not meaningful as the population was too small. For the spinal cord injured population however, this analysis did supply useful information. By means of Cattell's Salient Similarity index (CSSi) the pattern in factor loadings found in the spinal cord injured population was compared with the pattern found during the development of the SIP68. Judged by means of the CSSi the factor structure in this population did not differ significantly from the factor structure that was found during the development of the SIP68. This finding also can be considered to support the validity of the SIP68.

1.6.4. Conclusion

Based on the high to very high test-retest correlations, and the sufficient to good coefficients of internal consistency, it can be concluded that the SIP68 is a reliable instrument. Moreover, as far as the test-retest reliability at item level is concerned, it appeared not to make any difference whether the SIP68 is administered as a separate instrument or whether SIP68 scores are derived from the administration of the SIP136. This offers the possibility to calculate reliable SIP68 scores using data bases that contain the SIP136. The high test-retest correlations also can be interpreted as an indication of the stability of the concept measured. This interpretation is congruent with the expectations based on the model of health-related behavior that was developed in section 1.3. Another indication of the validity of the SIP68 as a measure of health-related functional status is found in the relations between the SIP68 and measures of related concepts (life satisfaction, level of functional independence, level of lesion). The correlations found are congruent with theoretical expectations, indicating (construct) validity. Two very different populations were used to study the reliability and validity of the SIP68. The instrument was studied in a population consisting of mainly relatively young men with a chronic and stable health problem (spinal cord injury) and a population containing predominantly older women suffering from (mostly slowly) progressive diseases (rheumatoid diseases). In both types of populations the instrument appeared to supply very useful information. Hence the preliminary conclusion of subsection 1.5.4 is confirmed and might be more firmly rephrased as: the SIP68 is a valid, reliable and more efficient generic alternative for the Sickness Impact Profile.
At this point in the SIP project, a number of the questions concerning the instrument had been answered: a theoretical model has been developed that provides a frame of reference to interpret SIP findings, the internal structure of the SIP136 was evaluated, based on findings in this evaluation, a short generic version of the instrument was developed, tested and found valid and reliable. The final topic that was studied in the SIP project was the ability of the instrument to detect relevant changes in the health-related functional status, in other words, the responsiveness of the SIP. As the SIP136 and the SIP68 might differ in their ability to demonstrate changes in functional status, in the final section of this chapter, findings on responsiveness of both versions of the instrument will be presented.

1.7. Responsiveness of the SIP136 and the SIP68

1.7.1. Introduction

Both the SIP136 and the SIP68 are designed to be used in cross-sectional as well as longitudinal studies. In a cross-sectional study design an instrument is intended to represent the level to which a certain concept is present. In a longitudinal design, however, the aim is not primarily or exclusively to assess the level present, but also to assess changes in that level of the concept aimed at. Therefore, apart from validity and reliability, the ability to detect changes (responsiveness) is an important

aspect of the quality of an instrument. It might even be argued that in the case of an instrument that is meant to be used in (longitudinal) evaluation studies, changes in the concept measured are part of the construct that is meant to be measured. Hence, the level to which an instrument is responsive (able to detect relevant changes) is part of the validity of an instrument. Considering the stable character of a SIP-score that can be derived from the health behavior model (section 1.3), changes are only to be expected when relatively large and drastic changes occur in the perception of the health status. This would mean that treatment that is directed at influencing the behavior or the behavioral expectations connected with the health status, is more likely to lead to a change in SIP score than treatment that is directly aimed at bio-physiologic health aspects. Changes in SIP scores following from traditional biomedical treatment that does not take into account that behavioral status is partly determined by social factors, are not expected to be large. In this case, only when the effect of biomedical treatment is clearly visible and obviously related to behavioral possibilities, are large SIP score changes to be expected. Several authors state that the SIP136 is responsive. Only few studies, however, explicitly evaluated the responsiveness of this instrument. Thus it was decided to devote the last part of the SIP project to answering the question whether the SIP136 and the SIP68 are able to detect relevant changes in the health-related functional status. To find an answer to this question, two other questions have to be answered first: a) against what criterion will the SIPs be judged, and b) what constitutes a relevant change in this criterion that should be registered by the instruments under study? As no generally accepted 'gold standard' exists for (changes in) the health-related functional status, and 'relevant changes' are not synonymous with statistically significant changes, responsiveness of these instruments can only be indirectly judged. The first step was to review relevant literature searching for information on changes detected by the SIP. Next the methods to study responsiveness would be inventarized and judged for their applicability. Finally available longitudinal SIP data would be used to study responsiveness of both the SIP136 and the SIP68.

1.7.2. Literature findings

The bases on which, in the literature, a judgement on the responsiveness of the SIP is grounded are very diverse. The first type of information found was the observation that SIP scores change. This change is usually accepted by the authors as a reliable and valid indication of a change in health-related functional status. As changes found are not related to any criterion, it is hard to judge whether these score changes indeed indicate a valid change in health-related functional status.
The second type of information is found in longitudinal studies where the SIP is used together with instruments measuring more or less related concepts. This kind of information offers the opportunity to relate SIP score changes to changes found by other instruments. In the majority of this kind of study, changes in SIP136 scores correspond with changes in the other instruments' scores as far as the direction of change is concerned. The most informative data was found in studies

that explicitly compare changes demonstrated by instruments measuring closely related concepts and the SIP. In a number of these studies the relation studied is expressed in the correlation between the SIP score changes and changes on the other instruments. This last type of information, however, is usually not directly reported by means of the actual correlation coefficients, but by an interpretation of the correlation found (e.g. 'strong correlations were found', 'significant correlations found' or even 'a correlation was found'). All studies reporting this kind of information, however, indicate that the SIP is adequate in showing changes. Only relatively few studies were found that do not report a change. This might be due to the so-called 'publication bias', studies that do find an effect of a treatment might have a bigger chance of being accepted for publication. This is a complicating factor when a conclusion is to be based on literature findings. Only one study was found that explicitly compares the responsiveness of the SIP136 with change shown by other 'general health' measures: the Arthritis Impact Measurement Scale (AIMS), the Functional Status Index (FSI), the Health Assessment Questionnaire (HAQ) and the Index of Well Being (IWB). It was concluded in this study that the SIP136, the HAQ and the IWB all 'are suitable to register change'. As none of these instruments is considered as a 'gold standard' for the others and no gold standard is present for 'general health', it cannot be said which one is the best.
In conclusion, based on the literature findings, the SIP136 demonstrates changes in the expected direction and (generally) in accordance with changes shown by other instruments, in many different types of populations and research settings. A definitive conclusion with respect to the responsiveness of the SIP136, however, cannot (yet) be drawn.

The fact that the information on responsiveness is rather vague characterizes the complex character of the responsiveness of instruments measuring abstract theoretical constructs like the SIP. There is no gold standard against which the changes that are found can be judged. Consequently it is not clear what change of SIP score corresponds with a relevant change in health-related functional status. Hence, a statistically significant change is usually accepted as 'real' change. Responsiveness, however, in the last few years has become generally recognized as a relevant psychometric topic to be judged in addition to to reliability and validity. Thus several methodological approaches that supply more informative data on responsiveness than mere statistical significance, were developed to capture this characteristic. However, no method or index has yet been generally accepted as the proper or ideal way to judge responsiveness. Roughly two types of approaches can be distinguished: correlation studies and the calculation of responsiveness indexes.
In correlation studies, the correlation coefficient is calculated between changes in the instrument under study and changes in a well established instrument measuring a related concept. The level of correlation coefficients reveals to what extent changes in one instrument are related to changes in the other. It is, however, arbitrary which threshold to use above which a correlation coefficient is sufficient to accept the claim for responsiveness. As the SIP68 and the SIP136 are measures of the same

concept, high correlations between change scores on both instruments would suggest equal responsiveness. The second method to judge responsiveness uses responsiveness indexes. Responsiveness indexes supply a standardized and dimensionless representation of the change demonstrated by a measure. It is expressed as the ratio of the average change found and an indicator of the accuracy of the assessment in question. Several responsiveness indexes were developed. As effect sizes are meant to quantify the amount of change measured, the standard deviation at base line (a cross-sectional figure) appears less accurate as denominator as it does not contain information on the accuracy of the instrument in detecting change over time. It seems more accurate to use the standard deviation of the change to standardize the amount of change found. In the present study into the responsiveness of both SIP versions, therefore, when an effect size is calculated, the second type (mean change/standard deviation of change) will be used.

1.7.3. Data analysis

Eight longitudinal data sets were available. Diagnoses in these sets were rheumatoid arthritis, ankylosing spondylitis, spinal cord injury, chronic back pain, back and neck complaints, cancer, stroke and Crohn's disease [25]. Three methods to assess responsibility were used to study the changes demonstrated by both SIP versions for every diagnostic group. First it was calculated whether changes registered by both instruments were statistically significant. In every sub-population where the SIP136 registers a statistically significant change, the SIP68 demonstrates a statistically significant change of approximately the same size in the same direction. The relation between changes detected by SIP136 and SIP68 was expressed in correlation coefficients. These coefficients ranged from 0.85 to 0.94 (median 0.91) for changes in total scores. Indicating a large amount of similarity between changes demonstrated by both SIPs. Correlations between changes in the SIP68 total score and the SIP136 dimension scores are less high. This difference in agreement will be due to the fact that the SIP68 aims at the total concept of health-related functional status, while both SIP136 dimension scores represent separate aspects of this overall concept.

A second method used was the comparison of changes registered by both SIPs in relation to changes found using measures of related concepts. In part of the available data sets five-point Likert scales are used to assess aspects of health or functioning (self-assessment of health, hindrance, physical functioning, psychologic functioning, severity of illness and pain). When a broad functional status measure like the Sickness Impact Profile is compared to a purely physiologic measure, very high correlations are not expected because the concept of functional status is far from identical to the concept measured by physiologic measures. A certain level of correlation, however, is expected, because on theoretical grounds one does expect a certain congruence between these two aspects of health. Between the available Likert scales and the SIPs on theoretical grounds a moderate to high correlation is expected, given the relatively close relation between the concepts measured. Correlation coefficients between significant changes detected by both SIP versions and by the available

self-assessments were calculated. In more than half the number of SIP-criterion combinations, no significant correlation was found. This might be partly explained by the difference in score range for the SIP136 and SIP68 on the one hand (0-100%, 0-68 points respectively) and the five-point scales (1-5 points) on the other. This difference in scales together with the relatively small changes registered by both SIPs might explain that relatively few significant or high correlations were found between SIP changes and self-assessment changes. However, significant correlations were found with every Likert scale in at least one of the sub-populations. These correlations all pointed in the expected direction: a positive change in SIP score coincides with a positive change on the self-assessment Likert scales. Highest correlations were found between both SIPs and self-assessment of health and physical functioning. Slightly less high coefficients were found with self-assessment of hindrance, psychological functioning and severity. These differences were congruent with the expectations. In general, when a significant correlation is found between SIP136 and a self-assessment, a correlation of similar height is found between the SIP68 and this scale. For both SIPs it can be concluded that the pattern of correlation coefficients confirms the construct validity of the changes measured as far as the direction and the differences in magnitude of correlation coefficients is concerned. As a third method of responsiveness assessment effect sizes are calculated for every significant difference in category, dimension or total score found between two assessments. According to the rule of thumb suggested by Cohen, all effect sizes found are small or at most moderate. In all populations except the spinal cord injured population, the SIP136 demonstrates a slightly larger effect size than the SIP68. These differences, however, are small. Using the method Tuley developed to compare the responsiveness of two instruments, the statistical significance of the difference in responsiveness between both instruments appeared not to be significant.

1.7.4. Conclusion

The first conclusion based on both the information found in literature and our own analyses is that both the SIP136 and the SIP68 probably are sensitive to changes in health-related functional status. Findings in empirical data suggest that the responsiveness of the SIP68 is to a very high degree similar to that of the original SIP136. The fact that changes are only found in populations that received relatively intensive treatment confirms the stable character of the concept of health-related functional status, derived from theoretical considerations in section 1.3. The generic character of both SIPs might also be due to the fact that only small changes were found. A generic measure is to be used in populations with all types and severities of health deviations. Therefore, only general potential functional consequences of a deviation in health status are checked by the instruments. Subtle functional nuances, characteristic for specific health problems, thus, might be 'overlooked' and the actual change is underestimated. Therefore, before final conclusions can be drawn concerning the responsiveness of the SIP68 and the SIP136, responsiveness of these instruments should be judged in relation to changes demonstrated by disease-specific measures of functional status.

A consideration at conceptual level is that only little is known about the mechanism that translates health problems into functional problems. The level of functional status measured by both SIPs is supposed to result from the health status. It might well be that when a change occurs in the physiological health status, this does not immediately result in a change of similar magnitude and direction in functional status. Before changes in measures of different definitions of health (functional status, bio-physiological health, psychological health) can be used to validate each other's responsiveness, the relation between these health definitions has to be studied in more detail. Moreover, when treatment is directed at changing bio-physiologic aspects of health, it is not to be expected that when these aspects are successfully treated, this would immediately result in a change in perceived health-related functional status. As was stated in section 1.3, it takes time for an individual to find out what behavioral consequences are connected with a health deviation. Moreover he or she has to reach an agreement with his or her social surroundings on the acceptable level of behavioral changes connected with the perception of the state of health.
Based on the findings in this study, it can be concluded that, in spite of the above considerations, the SIP136 and the SIP68 both appear to be responsive to changes in health-related functional status.

1.8. General conclusion

The focus of this project was the Sickness Impact Profile (SIP), a generic measure health-related functional status. A model was developed to give a theoretical basis to the interpretation of scores or score-changes of instruments measuring health-related functional status in general and the SIP in particular. The state of the art concerning the psychometric characteristics of the instrument was studied by means of a literature review. According to literature findings, the SIP is a valid and reliable measure of general health in many different situations. Information on the empirical evidence of the categorical structure, however, was not found. A negative remark frequently found was that the SIP is relatively long and thus not efficient. To obviate these drawbacks a short version was developed that has the same applicability as the original instrument. This instrument, the SIP68, was based on the internal structure of the list that was discovered in a series of principal components analyses. The resulting short instrument, thus, has an empirically validated internal structure. Evaluation of the SIP68 in populations with different diagnoses revealed that the validity and reliability of the SIP68 were similar to those of the original SIP.
From the model of health-related behavioral changes it was derived that health-related functional status does not directly and unambiguously result from the bio-medical health status. It results from the confrontation of views on the health status and the connected opinions on the 'true' or acceptable level of functioning. The impact of a health problem on the functional status is not given, but results from negotiations between all persons involved. In these negotiations the situation in terms of disease is redefined in terms of behavior expected from a person with a

given level and type of disease. This definition is based on the confrontation of implicit or explicit expectations, interpretations and prejudices of all involved about what behavioral consequences are 'normal' and 'justifiable' given the disease situation. This might explain part of the broad variety of functional consequences resulting from a given disease or level of disease. Once agreement is reached on what level of behavioral deviations is acceptable for a specific individual, the parties involved will not easily change their perception. Hence, the level of health-related functional status will have a relatively stable, recalcitrant or persistent character. This implies that a substantial change in functional status is only to be expected after a relatively large change in health status that is clearly and indisputably connected with functional consequences, or a clear change in the perception of what level of functional change is acceptable.
This negotiated character of health-related functional status implies that when the primary goal of treatment is not to cure the disease but to reduce or consolidate the functional consequences of health deviations (as in rehabilitation medicine or treatment of chronic, incurable health deviations), the perceptions of the situation by both the person under treatment and his surroundings should be taken into account. It might well be that the functional status can be manipulated by influencing perceptions and beliefs rather than influencing the actual physical situation. Traditionally, most health care interventions are primarily directed at the pathology. The implicit assumption is that when pathology is reduced, the functional status will improve. Following the health behavior model, this is only partly true. The empirical validity of this theoretical reasoning is supported by the fact that in this project both SIP versions only demonstrated significant score changes in populations with relatively large health problems, receiving intensive rehabilitation treatment. Rehabilitation treatment is usually not directed at curing disease but at reducing functional limitations that result from a chronic health deviation.

As far as the SIP is concerned the above considerations imply that a SIP score does not directly and unambiguously result from an underlying pathology. It results from the perception of what functional consequences are inevitably connected with the disease status. The same goes for a change in SIP score: this will not directly result from a change in disease status. A disease change will first be interpreted in terms of behavioral consequences by all involved, they will negotiate to reach an agreement on the 'true' health-related behavioral changes and this might result in a change in SIP score. Conversely, a change in SIP score does not necessarily stem from a corresponding change in disease status.
This health behavior model has implications for the application and interpretation of the SIP. Apart from pathology, subjective interpretations and beliefs concerning consequences of disease of the individual and his social surroundings, and the individual situational context are factors that influence a person when he is asked whether certain behavior is connected with his health status. The negotiated character of a SIP score implies that it does not measure a purely subjective or objective concept. The concept measured by the SIP instead can be characterized as intersubjective.

This intersubjective character entails that in longitudinal research the consequences of an intervention for the functional status will not be visible until negotiations have taken place and agreement is reached on the new acceptable level of functional status. Timing of post-intervention tests thus will have to be tuned according to the time needed to realize changes, negotiate and reach an agreement. When no change is found, this does not necessarily mean that the intervention did not have any effect. It might well be that the disease actually was effectively attacked, but the participants in the negotiations do not observe or believe that this will have any influence on the functional status. Thus the patient in question is still not expected to act as if he is healthy, and the sub-optimal functional status continues. This implies that when a treatment is primarily directed at bio-physiologic aspects of health (disease), an effect of this treatment will be registered more accurately by the SIP when within the treatment process attention is paid to the perception of functional or behavioral consequences of changes in health status.
On the other hand, when a change in SIP score is found, this does not necessarily imply that the disease status has changed. It also might indicate that views on what behavioral changes are inevitably connected with the disease status have changed. So, it might be that behavior actually changed (causing a change in SIP score) without any change or even with a reverse change in disease status. It might also be that specific behavior or behavioral changes that were believed to be caused or necessitated by the disease situation are no longer attributed to disease, causing a change in SIP score.
As it is to be expected that it takes time and effort before negotiations about the true level of health-related functional status result in an agreement between all participants, it is also to be expected that it will take time and effort to change this level. Changes in the functional status level (SIP score), therefore, are only to be expected some time after changes at disease level. Moreover, as all involved have to agree on the functional consequences of a disease situation, only clear and obvious changes in disease status will lead to SIP score changes. A change in SIP score is also more likely when functional aspects of health status are explicitly incorporated in the treatment process.
Based on the above considerations, in a cross-sectional research design the SIP is a good choice when respondents and their environment have had the chance to experience their functional possibilities and also have had the time to negotiate about the 'true' level of health-related functional status. In a longitudinal research design the SIP is the instrument of choice when treatment is directed not only or primarily at reducing disease but rather at influencing (enhancing) functional possibilities. These specifications lead to the conclusion that the SIP as a measure of general health is especially suited for the evaluation of treatment that is primarily directed at influencing the functional restrictions of people with incurable, disabling health deviations, although the instrument can also supply useful information concerning the functional status of more acute types of health deviations.
Once it is decided to use the SIP, in most situations the SIP68 is to be preferred over the SIP136. The SIP68 measures the same concept as the original instrument; it is also an indicator of the broad WHO definition of health. Moreover, the SIP68

meets the same high standards of reliability and validity as the original SIP, and it also has a generic character. As far as the responsiveness is concerned, both instruments appeared to be of similar quality. The major difference between the instruments is that the SIP68 contains only half the number of items of the SIP136 and uses a simplified procedure to calculate scores. This implies an important reduction in burden for both respondents and researchers in projects using the SIP68. However, when the instrument is meant to supply a detailed profile of an individuals functional status, the SIP136 is to be preferred above the more abstract SIP68.

The final conclusions of this project are the following

Based on an extensive literature study it can be concluded that the SIP136 is a reliable and valid measure of health-related functional status. The lack of empirical evidence of the validity of the categorical structure, and the relative length of the instrument, however, have led to the development of a generic alternative, the SIP68. This much shorter instrument also appeared to be a valid and reliable measure of the same concept as is measured by the SIP136. The ability of both instruments to measure changes in the concept aimed at cannot be expressed in a simple indicator of responsiveness. However, both instruments appear to correspond as to changes detected. It can therefore be concluded that they are equally responsive. Finally, the model of health-related behavior that was developed in this project offers a theoretical frame of reference that can be used to generate hypotheses on health-related functional status and changes in it, or to interpret findings of functional status measures.

Literature

1. Bergner M, RA Bobitt, WB Carter, BS Gilson. The Sickness Impact Profile: development and final revision of a health status measure. Med.Care 1981, 21: 787-805.
2. Luttik A, H Jacobs, LP de Witte. Een Nederlandse versie van de Sickness Impact Profile, Vakgroep Huisartsgeneeskunde, Rijksuniversiteit Utrecht, 1985.
3. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks: Sickness Impact Profile: the state of the art of a generic functional status measure, Soc.Sci.Med. 1992, 35: 1003-1014.
4. Parsons T. Definitions of Health and Illness in the light of American values and social structure, 1958; in: T. Parsons Social Structure and Personality, New York: Free Press, 1964.
5. Philipsen H. Afwezigheid wegens ziekte, Groningen, Wolters-Noordhoff, 1969.
6. Coe RM. Sociology of Medicine, New York, McGraw-Hill Book Company, 1978.
7. Mechanic D. Illness and social disability: some problems of analysis. Pacific Sociological Review 1959, 2:37-41.
8. Gerhardt U. Parsons, role theory, and health interaction, in: Scambler G. Sociological Theory and Medical Sociology, London Tavistock Publications 1987.
9. Thomas EJ. Problems of disability from the perspective of role theory. J. Health Hum. Behav. 1966, 7.
10. Strauss AL, J Corbin, S Fagerhaugh, BG Glaser, D Maines, B Suczek, CL Wiener. Chronical Illness and the quality of life. St. Louis, The C.V. Mosby Company, 1984.
11. Bergner M, RA Bobbit, S Kressel, WE Pollard, BS Gilson, JR Morris. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. Int. J. Health Serv. 1976, 6(3): 393-415.
12. Carmines EG, RA Zeller. Reliability and validity assessment. Sage, Beverly Hills 1979.
13. Cronbach LJ. Coefficient Alpha and the internal structure of a test. Psychometrica 1951, 16:297-334.
14. Nunnely JC. Psychometric theory. New York, McGraw, 1967.
15. Bergner M, RA Bobbit, WE Pollard, DP Martin, BS Gilson. The Sickness Impact Profile, validation of a health status measure. Med.Care 1976, 14: 57-67.
16. Kerlinger FN. Foundations of behavioral research. Holt, Reinehart and Winston, New York, 1975.
17. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen: The development of a short generic version of the Sickness Impact Profile, J.Clin.Epid. 1994, vol. 47, no 4: 407-418.
18. de Bruin AF, M Buys, LP de Witte, JPM Diederiks: The Sickness Impact Profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility, J. Clin. Epid. 1994, vol. 47, no.8: 863-871.
19. Rummel RJ. Applied Factor Analysis. Evanston: Northwest University Press, 1970.
20. Tabachnick BG, LS Fidel. Using multivariate statistics. New York Harper & Row, 1989.
21. Post MWM, AF de Bruin, LP de Witte, G Schrijvers: The SIP68: a measure of health-related functional status in rehabilitation medicine. submitted.
22. Bravo G, L Potvin. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J.Clin.Epidemiol. 1991, 44:381-390.
23. Jaccard P. Nouvelles recherches sur la distribution florale. Bul.Soc.Vaud.Sc.Nat. XLIV 1908, 163: 223-270.
24. Wade DT, C Collin. The Barthel ADL Index: a standard measure of physical disability? Int. Disab.Stud. 1988, 10:64-67.
25. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68, submitted.

Chapter 2

Sickness Impact Profile: The state of the art of a generic functional status measure

A.F. de Bruin[1], L.P. de Witte[1,2], F.C.J. Stevens[1], J.P.M. Diederiks[1,2].

[1] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands
[2] IRV, Hoensbroek, The Netherlands

This paper was published in Social Science and Medicine 1992, vol. 35, no. 8: 1008 - 1014.

Abstract

The Sickness Impact Profile (SIP) is a widely used health status measure, known to be valid and reliable. After the final develpoment and testing in 1978, however, in which several methodological aspects were investigated, no descriptions of research projects that systemetically evaluate the methodological and theoretical aspects of the instrument were found. In this article a review is presented of literature on the SIP. This review is the first step taken in a project that evaluates the SIP. The instrument appears to be a reliable instrument with sufficient content validity. It shows good correlations with other health status and functional status measures. Yet a number of questions about the SIP remain unanswered. Theoretical implications of the construct of sickness, the effect of age and gender on SIPscores, the construct validity judged by factor analysis, the responsiveness of the instrument, and the possibilities to use proxy-respondents or to shorten the list and to simplify the scoring procedure still have to be studied. If the instrument is to be used as an international standard measure of functional status, these topics should be thoroughly examined.

2.1. Introduction

The Sickness Impact Profile (SIP) is a behaviorally based self-report measure. It is an instrument that is used to gauge sickness-related dysfunction. It contains 136 items that are divided into 12 categories. In addition to category scores and the overall score, the instrument can be used to calculate a physical and a psychosocial dimension score. It is intended for use in measuring the outcomes of care in health surveys, in program planning, in policy formation and it is also used in monitoring patient progress. It is known as a valid and reliable measure of functional status. The SIP has been suggested as one of the three instruments to which research concerning the methodology of research on the quality of life and functional status measurement should be limited [1]. In an overview of existing measures in epidemiological research, it is stated that the SIP might well become a standard against which other measures are judged [2].
If research is to be limited to this instrument and if the SIP is really to become a standard measure, it is important to know what its methodological characteristics are and what research purposes it can and cannot serve. Is the SIP accurate in describing and delineating different populations? Does it efficiently mirror changes in functional status? What is the theoretical relevance of the SIP?
Judging by its reputation, the SIP's general reliability and validity have been established. However, in order to make a well informed choice with regard to using it in a specific research setting, more specific questions have to be answered. What has to be determined is the specific populations, administration types and research purposes to which the SIP can be applied. In this respect, the responsiveness of the SIP and the clinical relevance of its results also have to be established.
For practical reasons, the possibilities for shortening the instrument (eg, by selecting items or categories) should be explored. The result should be a more specific and compact, but still reliable and valid, measure. For both practical and theoretical reasons it is relevant to know whether the SIP can be used with proxy respondents to judge the functional status of a population or individuals.

The purpose of this article is to determine what information can be derived from publications on studies investigating or using the SIP, with respect to validity and applicability of the instrument and what information is found concerning the more specific aspects mentioned above. The method used in this article is a review of literature on the SIP. Most of the studies identified in this review do not consider the SIP as subject of study, but merely use information gathered with the instrument. Thus, because of insufficient information, meta-analysis of data on the SIP is not possible. In the present paper we gather information from published studies, on the basis of which we try to answer the above questions about the SIP.

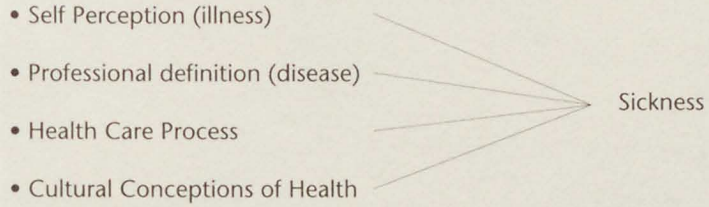
First, the SIP will be briefly introduced and the review method will be described. Next, the data found in the literature will be presented. Finally, conclusions with respect to validity, reliability and the aspects mentioned above will be drawn from these data.

2.2. The Sickness Impact Profile

The SIP is intended as a health status measure, to be used in health surveys, program planning, policy formation, and in monitoring patient progress in terms of sickness [3]. The developers of the SIP conceptualized sickness as 'changes in behavior associated with the carrying out of one's daily life activities' [3]. The developers intend to measure sickness impacts, which are defined as 'self reports of dysfunction, clinical reports of dysfunction, others' reports of dysfunction and tests of dysfunction' [3]. It can be concluded that the SIP does not measure the presence of medical, objective deviations (disease), nor does it aim to measure the subjective, experienced discomfort, or feeling state of the respondent (illness). While disease is objective and illness is subjective, sickness is an intersubjective construct. Sickness can be described as the changes in a person's behavior related to his or her (lack of) health. This means changes must be noticeable and recognizable by both the patient and his or her social surroundings. Therefore, there should be agreement between judgments from different judges of sickness. Assessment of sickness by the SIP is not solely based on individual judgments by patients or care providers, or on the (health) care process. The assessment of sickness is also influenced by intersubjective (cultural) aspects such as the conceptions of health and the patient's role in the population being studied (see fig. 2.1).

The SIP questionnaire consists of 136 items grouped into 12 categories. Each category represents a different aspect of daily functioning. Each item describes changes in behavior that are attributable to health status. Respondents are asked to mark only those items that are related to their health and that describe the respondents on a given day. Three categories (Ambulation, Mobility, Body Care and Movement) are aggregated into a physical dimension. Four other categories are aggregated into a psychosocial dimension (Social Interaction, Communication, Emotional Behavior, Alertness Behavior).

Figure 2.1. Model of Sickness (Impacts)



The five remaining categories (Eating, Work, Sleep and Rest, Household Management, and Recreation and Pastimes) are not aggregated). It takes about 20 to 30 minutes to administer the questionaire, and about 5 to 10 minutes to complete the scoring procedure. No special training is needed to administer and interpret the instrument and its results. Since the developers' last revision in 1981, the SIP has been used in a great variety of situations among different patient groups. Most of the groups are characterized by chronic or lengthy conditions (see Table 2.1). In most published studies the instrument appeared to fit the needs of the researcher. In several studies, however, the instrument was modified to fit a specific population or design (see 2.7.2).

In the United States, a Chicano-Spanish version of the SIP has been developed and described by Gilson [4]. Seven translations have been found in Europe: translations into British (a slightly modified American version called 'Functional Limitations Profile')[5C], Swedish [6], German [5], French (two translations), Danish [5A], Norwegian [7] and Dutch [5B]. Although two translations into French exist, no English publications have been found using the French SIP.

Table 2.1 Populations for which the SIP has been used (references between brackets)

Cardiovascular disorders	
hypertension	(10)
coronary/peripheral revascularization	(53)
cerebrovascular accident	(43,60)
cardiac arrest	(8)
coronary heart disease	(61)
heart problems	(62)
congestive heart failure	(63)
Neurologic disorders	
head injury	(19,20)
Alzheimer type dementia	(24)
spinal cord injury	(56,60,64)
neuro-muscular disease	(65)
intellectually impaired patients with Parkinson's disease	(66)
Internal disorders	
hyperthyroidism	(3,11)
cancer	(21,69,70)
end-stage renal disease	(49, 73)
abdominal complaints	(64)
chronic obstructive lung disease	(67)
chronic airflow limitation	(68)
Crohn's disease	(71)
diabetes	(72)
Pain and locomotory disorders	
rehabilitation	(3)
hip-replacement	(3)
low back pain	(15,17,37,59,77)
rheumatoid arthritis	(16,29,36,41,62, 64,74,75)
elderly undergoing knee-arthroplasty	(38)
chronic pain	(47,54,55)
ankylosing spondylitis	(71)
total joint replacement	(76)
chronic low back pain	(78)
Others	
medical and psychiatric problems	(22)
chronically or terminally ill elderly	(27,80)
variety of surgical diagnoses	(33)
women with urinary incontinence	(7)
injured workers	(40)
dental conditions	(48)
long term nursing home residents	(50)
speech pathology	(79)
chronic health problems	(79)
aged	(81)
psoriasis	(82)
morning sickness in pregnancy	(83)

The German translation has not yet been used in published studies [5]. With respect to the Danish and Norwegian SIP, the only publications found were written in Danish and Norwegian respectively, so that only the English abstracts could be read. For these reasons, no further information will be presented about the German French, Danish and Norwegian SIP versions.

2.3. Review method

The first step in collecting information was to define what information was needed to answer our questions. A scheme was developed containing the most important aspects of reliability, validity and applicability found in methodological literature. This scheme was used to identify and sort relevant information. The literature used in this paper was obtained from several sources. Firstly researchers in The Netherlands who had already done studies on the SIP were consulted. Secondly, a CD-rom machine readable citation base was used to screen MEDLINE using the words 'sickness', 'impact', and 'profile'. References were screened for possibly relevant papers that had not yet been selected. Because we also want to know in which countries the SIP is being used, we wrote to several well known researchers in the field of medical sociology in Germany, Canada, France, Belgium, Spain, Sweden and Denmark. We asked them whether they were familiar with any translations of the SIP in their countries, and if so, what their experiences were with using the instrument. Using this strategy, 120 articles describing aspects of the SIP or studies using the SIP were screened. We expect that we have collected the majority of studies published in which the SIP, its reliability, validity, modifications and translations are described. Evidence for this is the sharp decline in the number of new titles found as our literature search progressed. The items in the scheme which are used in the analysis of the literature, are described below. The data is found are presented, and, unless stated otherwise, no distinction is made between data coming from different translations.

2.4. Reliability

The best known measures of reliability are test-retest correlations and Cronbach's α (internal consistency). Data on these measures in relation to the SIP found in the literature are presented in table 2.2.

2.4.1. Test-retest reliability

The test-retest reliability appears to be good to very good for the overall SIP, its subdimensions and its categories. In a test-retest procedure, individual respondents did not mark exactly the same items every time the SIP was administered. This reflects individual daily variation in functional status. Nevertheless, the reproducibility of items marked, is satisfactory (r=0.45 to 0.60) [3,8].

Table 2.2 Reliability-coefficients of the SIP[1]

	overall SIP	dimensions	categories
Internal consistency			
- American SIP	0.94	0.91	0.60-0.90
- Swedish SIP	0.95		
- Spanish SIP	0.94	0.89-0.93	
- Dutch SIP	0.91	phys. 0.90 psy. 0.98	0.74 (0.45-0.87)
Test-retest			
- American SIP	0.86 (0.75-0.92)	phys. 0.90 psy. 0.79	0.50-0.56
- Swedish SIP	0.91	phys. 0.91 psy. 0.87	0.67-0.95
- Spanish SIP	0.85		
Interrater reliability			
r:			
- American SIP	0.92		0.72-0.92
kappa[2]			
- American SIP	0.87	phys. 0.80 psy. 0.65	

[1] In this table mean values found in literature are given.
[2] to assess kappa, SIP scores were divided into quartiles, since kappa is not applicable to continuous data.
sources: 3,15,16,6,26,28,29

2.4.2. Internal consistency

As can be seen in table 2.2, the internal consistency of the SIP shows that it has a high reliability. In all versions examined, the overall instrument, the psychosocial and physical dimension, and the separate categories appear to have sufficiently high α's.

2.5. Validity

In methodological literature five basic types of validity are found: content validity, criterion validity, construct validity, external validity and internal validity. Authors use different validity types, and no absolute agreement about the distinction between the types of validity is found in the literature. Although we are aware of

this fact, we nonetheless use these concepts to describe the validity of the SIP for reasons concerning the presentation of the data.

2.5.1. Content validity

Content validity concerns the extent to which the items in an instrument adequately represent the total property being measured (universe of content). Inspection of the items in an instrument by specialists or judging the item-selection procedure would confirm that the measure addresses the proper domain [9,10]. In developing the Sickness Impact Profile the aim was to collect a comprehensive catalogue of the possible impacts of sickness on behavior. SIP items ask about the three major aspects of health recommended by the World Health Organisation: physical, mental and social. Statements in the SIP items are expressed in behavioral terms, corresponding with the way in which patients generally describe their complaints [11]. In the development of the SIP, samples were gathered of statements describing sickness related changes in behavior from a broad range of individuals: patients, professionals, individuals caring for patients, and apparently healthy persons. The literature on health measurement was reviewed and additional statements were gathered. A series of field experiments reduced the number of items from the initial 1250 to 136. The items were clustered into categories.
In the final revision of the SIP, Bergner found that not all categories are required to account for variance among subjects in each subsample and on each criterion measure. Each category, however, appeared to be important in one or more situations. Bergner also observed a relatively low correlation among category scores, indicating minimal overlap between categories. The correlation of category scores to overall scores was higher, indicating the importance of each category for the whole instrument [3,8]. The fact that the SIP does not include a pain subscale is mentioned by several authors as a fundamental flaw in the instrument. Pain, however, is an aspect of the subjective health perception (illness), and therefore should not be included in an instrument measuring the intersubjective construct sickness. In this review of the literature about the SIP, no evidence against its content validity was found. The gathering of the items has been extensively described, and selection appears to have been performed with care. Most researchers using the SIP find a satisfactory face validity and do not question the content validity [eg. 12].

2.5.2. Criterion validity

Criterion validation refers to the technique of comparing scores on the instrument being studied with one or more external variables or criteria known or believed to measure the attribute being studied. A test is not valid unless it correlates significantly with that criterion [13,14]. As no gold standard for the assessment of function or health status is available, this kind of validity has to be assessed by using measures aimed at measuring constructs related, but not identical to, the phenomenon measured by the SIP. Table 2.3 presents data found on this characteristic.

Table 2.3. Coefficients of criterion validity of the SIP

	overall SIP	physical dimension	psycho-soc. dimension
Self assessment			
- sickness/gravity of symptoms	.54/.63/.46		
- (dys)function/hindrance	.52/.69/.52/.68	0.42	0.37
- health	0.51/0.37-0.621/0.50		
Clinical/doctor assessment of:			
- dysfunction	0.49/0.50		
- sickness	0.40		
- health	0.35/0.02-0.02[1]		
Health measures			
ADL index	0.46/0.74		0.42
National Health Insurance questions	0.61/0.55		
Quality of Well Being Scale	0.55		
Arthritis Impact Measurement Scale	0.87		
Functional Status Index	0.73		
Duke Health Profile	-0.70		
Health Assessment Questionnaire	0.78		
Index of Well Being	0.59		
Barthell Index	0.74		0.47
Nottingham Health Profile	0.55-0.71[1]		
General Health Rating Index	0.52		
Clinicians ARA ratings	0.30/0.66	0.36	0.02
Psoriasis disability index	0.45	0.40	0.35

[1] only over categories Sleep/Rest, Emotional Behavior, Ambulation, Mobility, Bodycare and Movement, Social Interaction.
sources: 3,8,10,16,32,36,39,61,62,77,82,84.

The correlations between criteria and the SIP provide evidence for the validity of the SIP and its translations. The SIP has a high positive correlation with self assessment of health status. However, correlations with clinical measures or doctor ratings are lower, though sufficient to assume clinical relevance of the SIP results. This less than perfect correlation between SIP scores and clinical or doctor ratings suggests that the SIP addresses a different, possibly broader construct than clinical measures or clinician ratings do. Correlations with more specific measures (like scales concerning activity of daily living and the official ratings of the American

Rheumatisme Association) are significant though not very high, also suggesting that the SIP measures a more broadly defined construct in which the specific features of these other instruments play a part. Correlations of the SIP with well known and often used general health or functional status measures like the Arthritis Impact Measurement Scale (AIMS), the Functional Status Index and the Health Assessment Questionnaire are high to very high, offering evidence of a good to very good criterion validity of the SIP. The validity of the Chicano SIP version, developed in a Chicano population in Washington State [4], appeared to be relatively poor when used in a Spanish speaking Mexican-American population [15]. These less positive results are possibly attributable to differences in education, culture and language between Chicano populations throughout the USA.
The general conclusion concerning the criterion validity of the SIP is that the overall SIP offers good criterion validity in a large number of situations and contexts. It offers information on a broader construct than clinical measures or specific aspect measures do.

2.5.3. construct validity

Construct validity is central to the measurement of abstract theoretical concepts. It concerns the extent to which a measure relates to other measures, in accordance with theoretically derived hypotheses about the concepts that are being measured [13]. Thus, while content validation is an inductive process, and criterion validity refers to a technique of validation, construct validation has a deductive nature.

Two types of construct validity have been widely accepted: convergent validity and discriminant validity. Convergent validity is demonstrated when a measure correlates with those measures to which it is related according to prior theory [9]. With regard to the SIP, this would imply that scores should be different for patient groups with diseases or severities of disease associated with different (levels of) disabilities or different (levels of) functional impairment. Discriminant validity means that the measure being studied should not be correlated with measures of constructs which according to prior theory are not related to the area of interest. This means that SIP scores should not correlate with features that are not influenced by the presence of the particular diagnoses, eg. health locus of control or educational level. In table 2.4, correlations are shown between the SIP and a large number of other measures.
From table 2.4 it can be concluded that correlations found are not very high. Neither psychological nor clinical or pain measures show very high correlations with SIP scores or with scores on parts of the SIP. What does appear from these data is that dimensions and categories each correlate on a different level with other measures. Differences usually occur as expected: physical measures correlate better with the physical dimension, the psychosocial dimension correlates better with psychological measures. The different associations of these dimension scores support the construct validity of these two dimensions.

Table 2.4. Coefficients of construct validity of the SIP

	tot.	phys.	psy.
Psychological measures			
Minesota Multiphasic Personality Inventory	.27-.53	.15-.32	.26-.64
Katz Adjustment Scale (emotional/psychological severity)	.46-.56		
Life Satisfaction Index Z	.16		.31
Philadelphia Geriatric Center Morale Scale	.19		.40
evidence of mental problems	.11	.03	.27
five item mental health scale	.57		
Profile of Mood States:			
- Vigor		-.44[1]	-.17[2]
- Depression/Dejection		-.24[1]	-.47[2]
Hospital Anxiety and Depression Questionnaire:			
- anxiety	.51	.43	.50
- depression	.62	.50	.66
Mood Adjective Checklist		.28	.64
Beck Depression Inventory	.60[5]		
Carroll Depression Rating Scale	.67	.44	.72
Clinical measures			
Glasgow Coma Scale	.26		
Halstead Impairment Index	.30		
number of past health problems	.25		
percent disabled	.15		
number of self-reported symptoms	.53		
RA Specific measures			
duration of sickness	.26	.36	.03
morning stiffness	.23	.15	.32
grip strength	.35	.33	.29
hematocrit	.29	.26	.12
sedimentation rate	.36	.44	.11
anatomic state	.17	.31	.01
Keitel index		.65[3]	.09[3]
Lansbury Articular Index		.56[3]	.21[3]
Richie Articular Index		.28[3]	.33[3]
c-reactive protein		.36[3]	.22[3]
Pain			
McGill Pain Questionnaire		.11[1]	.22[2]
rating worst pain last week		.21[1]	.34[2]
self-rated pain	.31	.33	.13
Body Symptom Scale (pain)[4]		-.43	-.39

	tot.	phys.	psy.
Others			
interviewer	.03(ns)		
employment status	.34		
activity rating	.71[5]		
time of day of interview	.21		
interview date	.02(ns)		
order of presentation	.05(ns)		
duration of interview	.36		
utilization of services	.19		
time achieved on treadmill	.63		
6 min. walking distance	-.64	-.72	-.50
forced vital capacity	.34		
nausea during pregnancy	.60		.42
age	ns		
years of schooling	.24		

[1] category Ambulation
[2] category Emotional Behavior
[3] categories Emotional Behavior, Social Interaction, Alertness, Communication
[4] higher score implies less discomfort
[5] categories Sleep-Rest, Emotional Behaviour, Home Management,Social Interaction, Work, Recreation and Pastimes
ns= not significant
sources: 10,16,17,20,21,22,41,47,77,68,83

From table 2.3 (criterion validity of the SIP) it also can be derived that dimensions and categories of the SIP have differential correlations with the criterion measures, indicating that these parts measure different aspects of the overall construct, and thus supporting the validity of the internal structure of the instrument. With regard to the discriminative or descriptive capacity, the SIP appears capable of describing and delineating groups of illnesses within varying test populations [3,8,12,16,17, 18,19,20]. Sullivan found that the category 'activation/deactivation' of the Mood Adjective Check List (MACL) correlates with all SIP categories. This suggests that the level of psychic energy is an important influence on the total SIP score.

The outcomes of factor analysis can be used to verify theoretical aspects of the construct being studied. This technique may reveal clusters of items which represent aspects of the construct being measured. The empirical relevance of categories and dimensions (the internal structure) might be verified by means of a factor analysis of the SIP. Information about a factor analysis of the SIP, however, is scarce: Sullivan reports on a factor analysis which was performed on the SIP scores and the MACL scores of 406 unselected cancer patients. It was found that all SIP categories loaded on the first factor and the MACL categories loaded on the second. This

means that the SIP measures other phenomena than the MACL does [6]. Greenwald performed a factor analysis on SIP-scores together with the Profile of Mood States (POMS) scores and the McGill Pain Questionnaire (MPQ) outcomes, gathered among seriously ill cancer patients. This analysis indicated that these three instruments each loaded on a different factor. The findings suggested that the SIP measures a construct in which mood states and pain are not incorporated [21], or at least, a construct that is not dominated by mood states or pain.
One study described a factor analysis on the SIP combined with some psychological measures. All SIP psychosocial categories loaded on one factor together with the other psychological measures. SIP physical categories loaded on another factor. In this study the conclusion was that the SIP measured at least two dimensions of health, one of which strongly related to depression [21]. From these findings, in which the SIP scores are confronted with different psychological measures, it can be concluded that SIP scores are influenced by mood status, but are not dominated by it. In other words: the SIP measures a concept in which psychological factors play a part. No study in which a factor analysis was performed on the SIP in order to investigate the internal structure of the instrument was found using a population other than the one used to develop the instrument. Apparently this aspect of the construct validity and the structure of the SIP has not been thoroughly investigated since it was developed.

2.5.4. Proxy respondents

The construct on which the SIP focusses is sickness. As mentioned before, sickness is an intersubjective construct. Because intersubjectivity is fundamental to the construct being studied, the use of proxy respondents should not merely be regarded as a convenient way of gathering data when patients cannot be asked for it themselves, but as an aspect of the construct validation of the SIP. Gilson et al. tested the use of proxy respondents on the SIP in the final development of the instrument. At best, they found a poor correlation between proxy SIPs and subject scores [23]. Krenz investigated this possibility by confronting the SIP scores provided by Alzheimer type dementia patients and SIP scores obtained from family members with scores derived from other more 'objective' measures. His data suggested that family member SIPs were valid measures of patients' functional status, and patient-SIPs were not, which led him to the conclusion that surrogate SIPs were also valid for non-demented patients [24]. Rothman found a correlation of 0.72 between the SIP physical dimension score from patients and primary caregivers, although proxies rated patients as slightly more impaired than patients rated themselves. The psycho-social dimension scores provided a correlation of 0.33. Rothman concluded that psychosocial proxy scores were not to be considered as valid substitutes for patient responses [24A].
In a study comparing the SIP scores of chronically or terminally ill patients and their caregivers, McCusker and Stoddard did not find sufficient agreement between the paired SIP-cores from patients and proxy respondents to warrant the use of surrogate SIPs at an individual level. At an aggregated level however, when proxy

respondents were closely involved family members, proxy SIPs appeared to be reliable and valid as surrogate scores for a group [25]. Zimmer used caretakers as substitute respondents on the SIP. Good levels of agreement were found between self-administered surrogates and patient-interview SIP scores [26]. When differences are found between proxy responses and patient responses in the assessment of sickness using the SIP, the difficult question that arises is whether patient responses should be the standard by which to judge the proxy responses or vice versa. Patient SIP's and SIP's completed by professional caregivers were compared by Goldsmith. He found high agreement between patients and caregivers about the absence of disability (items not marked). Agreement about the presence of disability, however, (items marked by both patient and careprovider) was not as high: patients marked twice as many items as caregivers [27]. When caregivers were informed about the responses of patients to the SIP, an increase was found both in recognised disabilities and in agreement between patient and doctor was found [27]. Although some researchers have investigated the possibility of obtaining SIP scores from proxy respondents, validity aspects of proxy responses have not yet been adequately tested [28,29]. Thus, they cannot be considered as interchangeable.

2.5.5. Internal validity

Internal validity concerns whether variations in the outcomes generated by the instrument being studied are due to variations in the construct, or to other influences that bias the score. Possible confounders are demographic variables and other characteristics of the test population, and the way the tests are administered. The influence of demographic variables like age, sex, education and employment status on SIP results has not been studied extensively. Deyo found that SIP scores may be correlated with age and sex [12,30]. Read found that age did not influence the score. Employment status and years of schooling did correlate with the SIP score, though not very highly [10,31]. (see table 2.4) Temkin found an indication that emotional status may have a major impact on SIP endorsement [20].

From table 2.1, it is clear that the SIP has been used to assess a great variety of diseases, which represent mostly chronic conditions. In all of these conditions the instrument appeared to be appropriate to measure health related functional status. In disease categories where respondents are cognitively impaired (eg. by Alzheimer type dementia), scores on the SIP cannot be taken at face value. For these cases the use of proxy respondents is a possible solution (see 5.4). Most published studies found no evidence of selection bias, although some of the research populations were not representative of the total population from which they were drawn. Therefore it seems justified to consider that SIP scores obtained from patients in most disease categories present good, useful data.

As far as patients' reactions to the SIP are concerned, Deyo found that only 2% of the reactions were negative (71% positive, 27% neutral). Typical comments were as follows: 'did not understand some words', 'too long', 'a lot did not apply to

me'. Filling out the SIP did not cause distress or discomfort in most respondents (97%). Comments were as follows: 'a little negative', 'too many questions and very stressful' [32]. The way and order in which the SIP was administered differs from project to project. In most cases, it was administered as part of an interview by a partially trained interviewer. In some cases, however, respondents filled in the list by themselves after receiving instruction. In a few projects, the SIP was delivered by mail and filled in by the respondent while no administrator or interviewer was present. In one case, the SIP first was administered in a personal interview by an interviewer, followed by a second SIP by telephone. These procedures all provided reliable and valid data. Bergner compared three types of administration; although some differences were noted, no administration type consistently displayed stronger relationships to criterion variables than any other did. For mail-delivered SIPs, internal consistency and correlation with criteria were somewhat lower, though still sufficient [3,8,33,34]. Mail- delivered SIPs did not produce results identical to interviewer and self-administered SIPs or SIPs administered by an interviewer by telephone. Still, mail delivered SIPs delivered reliable data [33]. Read found that the order of presentation and different interviewers did not influence scores [10] (see table 2.4). In short, the way and order of administering SIP and diagnoses of responders (possible sources of bias), do not seem to harm the internal validity of the instrument.

2.5.6. External validity

In order to study external validity, findings in one project should be tested for their consistency with results from projects studying other populations. Findings from more or less similar populations with the same diagnoses should correspond with each other. Results from different diagnoses should be consistent with similarities and differences in disease and population characteristics. Very few data have been found on the generalizability of SIP findings over populations with different diagnoses. Sullivan is the only author found in this review who explicitly mentions external validity. She concludes that: 'good evidence of generalizability (was) found with the SIP in cancer, rheumatic disorders, benign chronic pain, spinal cord injuries, chronic renal insufficiency, obesity, diabetes and angina pectoris' [6]. From this paper it is not clear how she reaches this conclusion and on what aspect or factors this generalizability was found.
Three factors may influence the external validity: testing, selection and reactive factors [35]. These influences may lead to findings only applying to the population and situation of a specific project. No evidence for this kind of bias was described in the literature reviewed when the SIP was used. Findings regarding the external validity of the SIP thus are not totally convincing. No projects searching for generalizable findings are described.

2.6. Responsiveness

'Responsiveness' or 'sensitivity to change' refers to the likelihood of detecting a clinically important change or treatment effect. While reliability and validity are sufficient to conclude whether an instrument is useful for descriptive or predictive purposes, for an evaluative instrument like the SIP, we must also have information about the instrument's responsiveness.
This feature is of specific interest because responsiveness may affect the statistical power of a (clinical) trial: the more responsive an instrument, the greater the possibility of detecting changes or differences when they are present [16,32,36]. Only in a few, more recent studies has the responsiveness of the SIP been assessed. A problem in assessing responsiveness is that a statistically significant change in test results over time need not be similar to a clinically important change. Several methods to assess and to quantify responsiveness have been suggested, but as yet no standard procedure is available [32]. In order to find a test that has the proper level of responsiveness needed in an evaluative project, it is necessary to know the clinical meaning of changes found with the instrument.

The SIP is meant to be used in longitudinal designs. In most studies, however, it was evaluated as a descriptive or predictive instrument to discriminate between or within groups in cross sectional research designs. A discriminative index should have a good reliability and should detect differences between individuals at one moment in time. An evaluative instrument should detect clinically important changes over time. Little is known about the ability of SIP to detect those changes. MacKenzie found that the SIP was mainly responsive to (large) changes for the worse in respondents with initially mild dysfunctions in the psycho-social dimension. In this study, functional improvement, either psycho-social or physical, and deterioration in physical functioning, were not accurately detected [34]. Deyo studied the responsiveness of the SIP in three different ways [37]. He did not find a clear answer to the question of responsiveness. However, he concluded from his findings that a brief disease-specific scale might have advantages in detecting change over a lengthy generic instrument like the SIP [37]. Follick, Nielson, Kubo, Hunskaar, Ahlmèn and Caradoc-Davies on the contrary, did find changes over time, using the SIP [7,17,38,39,40]. The changes found by these authors in small specific populations were mostly in accordance with changes in scores from the same respondents on other measures. However, not all changes found, however, were statistically significant.

In Sweden, the SIP was found to be sensitive to 1 year pre- and post-treatment changes, both for improvements and deteriorations. Changes in the SIP physical dimension corresponded more closely with changes in clinical findings than did changes in the SIP psycho-social dimension [41]. The Dutch SIP version was tested on responsiveness in a small trial, and appeared not to be sensitive to daily functional changes: incidental diary fluctuations were not reflected in the SIP scores, but

fluctuations over a longer period (3 weeks) were detected by the SIP [42]. Mulders, however, found that the Dutch SIP overall score and scores in several categories did significantly reflect a change in function in stroke patients [43]. Bergner stated that 'the clinical significance of the SIP score differences as with many tests is not firmly known' [44], prompting the question of clinical relevance of SIP scores or score changes. No general standard has been defined concerning the clinical implications of SIP scores. In other words, it is not known which SIP scores indicate the necessity of treatment and which scores are acceptable or normal. This question is most relevant if the SIP is used in clinical practice to evaluate individual patient progress. Although the instrument has been used this way, no answer was found to this question.

To summarize, a definite conclusion about the responsiveness of the SIP cannot yet be drawn. No generally accepted method for assessing responsiveness has been found and no criterion is generally accepted. Findings suggest that the instrument is not sensitive to small, daily changes in a patient's situation, but changes occurring over a longer period of time seem to be mirrored in SIP scores. Further investigation into this matter is needed.

2.7. Feasibility

2.7.1. Scoring procedure

To increase discrimination, precision, and sensitivity in accounting for variance, the instrument was scaled by its designers. Items were weighted for their relative importance by 25 health professionals and students. A second scaling procedure was performed by 108 members of a pre-paid group health plan. This resulted in weights representing the relative impact of the phenomenon described in the item on the health related functional status. These weights were used to calculate the scores for the categories, dimensions and the overall instrument [3,8,28,29]. The weights used in the official scoring procedure are social preference weights which are meant to reflect the relative importance of the items. In the total or dimension score the relatively small differences in weights could be 'drowned' by the large number of items, and therefore hardly influence the ordering of individuals.

According to methodological literature, nominal weighting, as used in this case, has its greatest effect on the ordering of individuals in composite scores when there is considerable variation in the weights, when there is little inter-correlation among the components, and when there is a small number of components or items [45,46]. The SIP has a relatively large total number of items, and has been developed with consideration for internal consistency (high item inter-correlations). If the weighting is to be done on the basis of item reliability or item validity, the variation among weights will be minor. As stated above, the differences between the item weights in the SIP are too small to have an important impact on the total

or dimension score. Therefore the rank order of persons in a distribution of weighted scores will be almost identical to the rank order of the distribution of unweighted scores. An alternative is not to use weights but to simply add the number of items checked. In a sense this procedure uses weights of 1 for every item. This is the simplest procedure [45,46]. To determine whether this simple scoring procedure of the SIP is as reliable and valid as weighted scoring, correlations between traditionally derived and dichotomous SIP scores should be examined. Skipping the weighting procedure would simplify the scoring procedure and thus improve the efficiency of the instrument. Moreover, it would make examination of the instrument possible, using principles of the item response theory, like Rasch analysis. Validity and reliability aspects are judged on the basis of classical test theory which essentially is only concerned with reliability. If dichotomous SIP categories appear to fit the Rasch model, (clinical) interpretation of individual SIP category scores would be possible, independent of the specific sample from which the respondent is drawn. This would improve the clinical relevance, applicability and acceptability of the instrument.

2.7.2. Modifications of the SIP

The categorical structure of the SIP permits measurement of well defined areas of activity or behavior. Several researchers have used parts of the SIP in their studies [34,47,48,49,50,51,52,53]. It is not always clear from the description of the study whether the total instrument was administerd and only a part of the total SIP was used in the analysis, or whether authors only administered those parts of the instrument they thought relevant for their study. Others (slightly) modified the instrument to adjust it to their specific purposes [6,17,19,47,54,55,56]. Johnson et al. changed the statements into questions to facilitate the instrument's use in an interview. A high correlation between this and the original format was established: r=0.94 [57]. It is not clearly described in every study whether the modifications or shortened versions had consequences for validity, reliability or the weighting and scoring procedure. In general, the SIP seems to be robust enough to serve the respective authors' purposes accurately even if it was modified or shortened.
Roland extracted a 24-item SIP version from the British SIP (the Functional Limitations Profile): the Roland Scale, designed for acute low back pain patients. Judged by correlations with physical measures of disease severity and with the overall SIP, this scale appears at least as valid as lengthier scales [58,59]. Sullivan and her team developed and tested several SIP modifications [6]. A shortened SIP version of eight statistically derived items optimally discriminated between respondents and non-respondents to therapy in chronic pain patients [54].
A shortened, specific version of the SIP appeared applicable across tumor types and sites [41]. The SIP was also used as a source for items concerning functional status in the development of a questionnaire for spinal cord injury patients [56]. A short rheumatoid arthritis specific SIP version is underway [41]. Most of these modifications appear to discriminate between health levels, evaluate patients responses to treatment and predict future states [6,41,54,56].

2.8. Conclusions

According to the literature reviewed, the SIP generally appeared to be a reliable and valid measure of functional status. However, some qualifications must be provided here.
Positive data were found on the reliability of the SIP. Judged by test-retest procedures and Cronbach's α the SIP has good reliability. The content validity also appears to be acceptable: no fundamentally critical remarks on the content of the instrument were found. By correlating other 'health-status' or 'functional-status' instruments with the SIP, the criterion validity was assessed and found to be good. As to the internal validity of the SIP, the influences of age, sex, schooling and employment status on SIP results have not been studied extensively. As to other aspects of internal validity, the literature reviewed shows these to be of an acceptable level within one disease category. The instrument appears to be applicable in populations that suffer from very diverse diseases. Most studies, however, focus on chronic or lengthy conditions. Diseases in which the cognitive capacities of the respondents are affected appear to limit the application of the SIP to the use of proxy respondents on an aggregated group level. Mail delivered SIPs as well as interviewer and self-administered SIPs or SIPs administered by telephone obtain reliable data. The instrument appeared sufficiently robust to allow slight adaptation of the content to fit the requirements of specific researchers or populations. Sub-dimensions or separate categories can be used separately to measure more specific aspects of health or functional status.

More critical remarks ought to be made on some other topics like the construct validity. From the literature, factor analysis appears to be one of the most suitable procedures to test construct validity of a measurement instrument. Hardly any data were found on this subject: no extensive description of factor analysis was found. The external validity could not sufficiently be supported by data from this review. Papers focussing on generalizability have not been found. With respect to the responsiveness and clinical relevance of the instrument, no definite conclusions can be drawn, because data on these subjects were sparse. The same can be said about the possibility of using proxy respondents to fill in the SIP. This has not yet been sufficiently explored, though some promising initial data on this subject were found. Though several successful attempts have also been made to shorten the list of 136 items for use in specific populations, no generally applicable short SIP version was found. Finally, for practical and statistical reasons it seems desirable to simplify the scoring procedure of the instrument.

From the data found in this review it can be determined that the SIP probably accurately describes and delineates different populations. Also, it appeares to be able to differentiate between different levels of functional status within a diagnosis group. Whether the SIP efficiently mirrors clinically relevant changes in functional status is not yet firmly known. Therefore, until more is known about the responsiveness, the instrument seems better fit for cross-sectional designs than for longitudinal designs.

No publications have been found about the theoretical relevance of SIP scores (eg., what its contribution might be to theory on the chronically ill).
In short, the Sickness Impact Profile appears to be a generally good, valid and reliable instrument to describe functional status, though some questions about methodological characteristics still have to be answered. In our further research concerning the SIP we will focus on these questions.

Literature

1. Spitzer WO. Keynote address: State of science 1986: quality of life and functional status as target variables for research. J.Chron.Dis. 1987, 40(2).
2. McDowell I, N Newell. Measuring Health: A guide to rating scales and questionnaires. Oxford University Press, New York, Oxford, 1987.
3. Bergner M, RA Bobbit, S Kressel, WE Pollard, BS Gilson, JR Morris. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. Int.J.Health Serv. 1976, 6(3).
4. Gilson BS, D Erickson, CT Chavez, RA Bobbit, M Bergner, WB Carter. A Chicano version of the Sickness Impact Profile. A health care evaluation instrument crosses a linguistic barrier. Cult.Med.Psychiatry 1980, 4: 137-150.
5. Potthoff P. 1989, personal communication.

5A. Folker H. Sickness Impact Profile-SIP, en metode til vurdering af helbred. Ugeskr. for Laeger 1987, 149(3).

5B. Jacobs HM, A Luttik, FWMM Touw-Otten, RA de Melker. The Sickness Impac Profile; results of an evaluation study of the Dutch version. Ned.Tijdschr.Geneesk. 1990, 134(40): 1950-1954.

5C. Patrick D. Standardization of comparative health status measures: using scales developed in America in an English speeking country. Survey research methods: Third biennial conference, MD.: US Dept. of health and human services, pub. no.(PHS) 81-3268, 1981 216-220.

6. Sullivan M. The Sickness Impact Profile: an instrument for overall health assessment; a basic evaluation. TGO/JDR 1988, 13: 167-169.

6A. Bergner M, RA Bobbit, WE Pollard, DP Martin, BS Gilson The Sickness Impact Profile, validation of a health status measure. Medical Care, 1976, 14(1).

7. Hunskaar S, A Vinsnes. The quality of life in women with urinary incontinence as measured by the Sickness Impact Profile. J.Am.Ger.Soc. 1991, 39: 378-382.

9. Meerling. Methoden en technieken van psychologisch onderzoek. Boom, Meppel/ Amsterdam, 1981.

10. Read JL, RJ Quinn, MA Hoefer. Measuring overall health: an evaluation of three important approaches. J.Chron.Dis., 1987, 40: 7S-21S.

11. Rockey PH, RJ Griep. Behavioral dysfunction in hypotheroidism, improvement with treatement. Arch.Int.Med., 1980, 5: 247.

12. Deyo RA, TS Inui, J Leininger, S Overman. Physical and psychosocial function in rheumathoid arthritis. Clinical use of a self-administrered health status instrument. Arch.Intern.Med., 1982, 142.

13. Carmines EG, RA Zeller. Reliability and validity assessment. Sage Publications. Beverly Hills/ London, 1979.

14. Kerlinger FN. Foundationns of behavioral research. Holt, Rinehart and Winston inc., 1975.

15. Deyo RA. Pitfalls in measuring the health status of Mexican Americans comparative validity of the English and Spanish SIP. Am.J.Publ.Health, 1984, 74(6): 569-73.

16. Deyo RA, TS Inui, J Leininger, S Overman. Measuring functional outcomes in chronic disease: A comparison of traditional scale and a self-administered health status questionnaire in patients with rheumatoid arthritis. Med.Care 1983, 21(2).

17. Follick MJ, TW Smith, DK Ahern. The Sickness Impact Profile: a global measure of disability in chronic low back pain. Pain, 1985, 21: 67.

18. Keller C. Psychological and physical variables as predictors of coping strategies. Percept. Mot. Skills. 1988, 67(1): 95-100.

19. Temkin N, A McLean, S Dikmen, J Gale, M Bergner, MJ Almes. Development and evaluation of modifications to the Sickness Impact Profile for head injury. J.Clin.Epidemiol., 1988, 41: 47-57.

20. Temkin NR, S Dikmen, J Machamer, A McLean. General versus disease specific measures. Further work on the Sickness Impact Profile for head injury. Med.Care 1989 27: Supplement.

21. Greenwald HP. The specificity of quality of life measures among the seriously ill. Med.Care 1987, 25: 642-51.

22. Blair Brooks W, JS Jordan, GW Divine, KS Smith, FA Neelon. The impact of psychologic factors on measurement of functional status: assessment of the Sickness Impact Profile. Med.Care 1990, 28.

23. Gilson BS, M Bergner, RA Bobbit, WB Carter. The sickness impact profile: final development and testing 1975-1978. Seattle, Washington: University of Washington, Department of Health Services, 1979.
24. Krenz C, EB Larson, DM Buchner, CG Canfield. Charactering patients dysfunction in Alzheimers-type dementia. Med.Care 1988, 29: 453-61.
24A. Rothman ML, SC Hedric, KA Bulcroft, DH Hickam, LZ Rubenstein. The validity of proxy-generated scores as measure of patient health status. Med.Care 1991, 29: 115-124.
25. McCusker J, A Stoddard. Use of a surrogate for the Sickness Impact Profile. Med.Care 1984, 22: 789.
26. Zimmer JG. A randomized control study of a home health care team. A.J.P.H., 1985, 75(2).
27. Goldsmith G, M Brodwick. Assessing the functional status of older patients with chronic illness. Fam.Med. 1989 21: 38-41.
28. Bergner M, RM Kaplan, JE Ware. Evaluating health-measures, commentary: measuring overall health, an evaluation of three important approaches. J.Chron.Dis. 1987, 20: 23S-26S.
29. Bergner M. et al. Health status measures: an overview and guide for selection. Ann.Rev. Pub.Health, 1987, 8: 191-210.
30. Deyo RA, TS Inui. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Serv.Res. 1984, 19(3): 277-89.
31. Hart LG, RW Evans. The functional status of ESRD patients as measured by the SIP. J.Chron.Dis., 1987, 40: 117S-130S.
32. Deyo RA. et.al. Barriers to the use of health status measures in clinical investigation, patient care and policy research. Med.Care, 1989, 27: S254-S68
33. Bergner M, RA Bobbit, WB Carter, BS Gilson. The Sickness Impact Profile: development and final revision of a health status measure. Med.Care 1981, 21(8).
34. MacKenzie CR, ME Charlson, D DiGioia, K Kelley. Can the Sickness Impact Profile measure change? An example of scale assessment. J.Chron.Dis., 1986, 39(6): 429-238.
35. Cook TD. et.al. Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago, Rand McNally, 1979.
36. Kirshner B, G Guyatt. A methodological framework for assessing health indices. J.Chron. Dis., 1985, 38: 27-36.
37. Deyo RA, RM Centor. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J.Chron.Dis. 1986, 39(11): 897-906.
38. Nielson WR, AW Gelb, JE Casey, FJ Penny, RN Merchant, PH Manninen. Long-term cognitive and social sequelae of general versus regional anesthesia during arthroplasty in the elderly. Anesthesiology, 1990, 73: 1103-1109.
39. Ahlmèn EM, CB Bengtsson, BM Sullivan, A Bjelle. A comparison of overall health between patients with rheumatoid arthritis and a population with and without reumatoid arthritis. Scand.J.Rheumatol. 1990, 19: 413-421.
40. Caradoc-Davies TH, BD Wilson, JG Anson. The rehabiltation of injured workers in New Zealand: a pilot study. N.Z.Med.J., 1990, 103: 179-182.
41. Sullivan M, M Ahlmén, LE Augustinsson, C Bengtsson, I Branehög, C Lundqvist, L Sjöström, J Cohen. Health Status Assessment in Rheumatoid Arthritis. Part I: Further work on the validity of the Sickness Impact Profile. J.Rheumatology 1990, 17: 439-47.
42. Witte LP de, H Philipsen, M vd Maegdenbergh. Stabiliteit van SIP-scores in een periode van 3 weken bij RA-patiënten. internal publication, University of Limburg, Maastricht, 1987.
43. Mulders AHM, LP de Witte, JPM Diederiks. Evaluation of a rehabilitation after-care programme for stroke patients. J.Rehab.Sciences 1989 2(4).
44. Bergner L, AP Hallstrom, M Bergner, MS Eisenberg, LA Cobb. Health status of survivors of cardiac arrest and of myocardial infarction controls. A.J.P.H., 1985, 75(11): 1321-23.
45. Ghiselli EE. Measurement theory for the behavioral sciences. W.H. Freeman and Company, San Francisco, 1981.
46. Guilford JP. Fundamental statistics in psychology and education. McGraw-Hill International Book Company, 1981.
47. Watt-Watson JH, JE Graydon. Sickness Impact Profile: A measure of dysfunction with Chronic Pain Patients. J.Pain Symp.Management, 1989 4(3).

48. Reisine ST, J Fertig, J Weber, S Leder. Impact of dental conditions on patient's quality of life. Com.Dent.Oral.Epidemiol. 1989, 17: 7-10.
49. Julius M, VM Hawthorne, P Carpentier-Alting, J Kneisley, RA Wolfe, FK Port. Independence in activities of daily living for end-stage renal disease patients: biomedical and demographic correlates. Am.J.Kidney Dis. 1989, 13(1): 61-69.
50. Abrams JP, HF Wallach, S Divens. Behavioral improvement in long term geriatric patients during an age integrated psychosocial rehabilitation program. J.Am.Geriatric Soc. 1979, 27: 218.
51. Bergner M. Measurement of health status. Med.Care, 1985, 23: 696-704.
52. Hall J, E Fisher, D Killer. Measurement of outcomes of general practice: comparison of three health status measures. Fam.Pract. 1987, 4(2): 117-22.
53. King KB. Patient management of pain medication after cardiac surgery. Nurs.Res. 1987 36(3): 145-50.
54. Augustinsson LE, L Sullivan, M Sullivan. Physical, psychologic and social function in chronic pain patients after epidural spinal electrical stimulation. Spine 1986 11(2): 111-9.
55. Augustisson LE, L Sullivan, M Sullivan. Chronic pain in functional neurosurgery: function and mood in various diagnostic groups with reference to epidural spinal electrical stimulation. Schmerz, Pain, Douleur 1989, 10: 30-40.
56. Siösteen A. Adjustment to spinal cord injury. A clinical and experimental study on quality of life, sexuality and fertility. Doctoral thesis 1989, Gothenburg University, Göteborg, Sweden.
57. Johnson J, K King, R Murray. Measuring the impact of sickness on functions of radiation therapy patients. Oncol.Nsg.Forum, 1983, 10: 36-39.
58. Deyo RA. Comparative validity of the SIP and shorter scales for functional assessment in low back pain. Spine 1986 11(9): 951-4.
59. Roland M, R Morris. A study of the natural history of back pain: Part 1: development of a reliable and sensitive measure of disability in low-back pain. Spine, 1983, 8: 141-144.
60. Witte LP de, H Jacobs, F vd Horst, A Luttik, J Joosten, H Philipsen. De waarde van de Sickness Impact Profile als maat voor het funktioneren van patiënten. Gezondheidszorg en Samenleving, 1987 8(2).
61. Salek MS, MJ Vandenburg. Measuring the quality of life in angina pectoris. TGO/JDR, 1988, 13(5): 186.
62. Ott CR, ES Sivarajan, KM Newton, MJ Almes, RA Bruce, M Bergner, BS Gilson. A controlled randomized study of early cardiac rehabilitation: The SIP as an assessment tool. Heart-Lung, 1983, 12(2): 162-70.
63. Tandon PK, H Stander, RP Schwarz jr. Analysis of quality of life data from a randomized, placebo-controlled heart-failure trial. J.Clin.Epid. 1989, 42: 955-962.
64. Witte LP de, H Jacobs, A Luttik. The Sickness Impact Profile: A Dutch Version. Poster.
65. Terpstra Sj. Leven met een spierziekte. internal publication, University of Limburg, 1990.
66. Sano M, Y Stern, K Marder, R Mayeux. A controlled trial of piracetam in intellectually impaired patients with Parkinson's disease. Movem.Disorders 1990, 5: 230-234.
67. Keller C. Predicting the performance of daily activities of patients with chronic lung disease. Percept.Mot.Skills, 1986, 63: 647-651.
68. Jones PW, CM Baveystock, P Littlejohns. Relationsips between general health measures with the Sickness Impact Profile and respiratory symptoms, psychological measures, and mood in patients with chronic airflow limitation. Am.Rev.Respir.Dis. 1989 140: 1538-1543.
69. Courtens AM. Longitudinal study of continuity of care for cancer patients. Internal publication, University of Limburg, Maastricht, 1988.
70. Sugerbaker PH, I Barkofsky, SA Rosenberg, FJ Gianola. Quality of life assessments of patients in extremity sarcoma clinical trials. Surgery, 1982, 91.
71. Janssen M, H Philipsen. Dynamiek van sociale netwerken van chronisch zieken en gehandicapten. Tijdschrift van Sociale Gezondheidszorg 1986, 64: 740-741.
72. Mitchell BD, MP Stern, SM Haffner, HP Hazuda, JK Patterson. Functional impairment in Mexican Americans and non-Hispanic Whites with diabetes. J.Clin.Epid. 1990, 43: 319-327.
73. Deniston OL, FA Luscombe, DP Buesching, RE Richner, BS Spinowitz. Effect of long-term epoetin beta therapy on the quality of life of hemodialysis patients. ASAIO Transactions 1990 36: M157-M160.

74. Ahlmén M, M Sullivan, A Bjelle. Team versus non-team outpatient care in rheumatoid arthritis. A comprehensive outcome evaluation including an overall health measure. Arthritis Rheum, 1988, 31: 471-79.
75. Sullivan M, M Ahlmén, B Archenholtz, G Svensson. Measuring health in rheumatic disorders by means of a Swedish version of the SIP. Results from a population study. Scand.J.Rheumatol. 1986, 15(2): 193-200.
76. Liang MH, K Cullen, M Larson. In search of a more perfect mousetrap. (Health status or quality of life instrument). J.Rheumatol. 1982, 9: 5.
77. Deyo RA, AK Diehl. Measuring physical and psychosocial function in patients with low back pain. Spine 1983, 8(6): 653-42.
78. Keijzers JFEM, L Bouter, H Philipsen. A back school in The Netherlands: evaluating the results. Pat.Educ.Couns. 1989, 14: 31-44.
79. Pollard WE, RA Bobbit, M Bergner, DP Martin, BS Gilson. The Sickness Impact Profile: reliability of a health status measure. Med.Care, 1976, 14: 146-155.
80. Yergan J, J LoGerfo, S Shortell, M Bergner, P Diehr, W Richardson. Health status as a measure of need for medical care: A critique. Medic.Care, 1981, 12: Supplement.
81. Bess FH, MJ Lichtenstein, SA Logan, MC Burger, E Nelson. Hearing impairment as a determinant of function in the elderly. J.Am.Ger.Soc. 1989, 37:123-128.
82. Finlay AY, GK Khan, DK Luscombe, MS Salek. Validation of Sickness Impact Profile and Psoriasis Disability Index in psoriasis. Br.J.Derm. 1990, 123: 751-756.
83. Hyde E. Acupressure therapy for morning sickness. J.Nurse-Midwifery 1989 34(4): 171-178.
84. Parkerson GR, WE Broadhead, CKJ Tse. The Duke Health Profile, a 17-item measure of health and dysfunction. Med.Care. 1990, 28(11).

Chapter 3

The development of a short generic version of the sickness impact profile

A.F. de Bruin[1], J.P.M. Diederiks[1,2], L.P. de Witte[1,2], F.C.J. Stevens[1], H. Philipsen[1].

[1] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands
[2] IRV, Hoensbroek, The Netherlands

This paper was published in the Journal of Clinical Epidemiology 1994, vol. 47, no.4: 407 - 418.

Abstract

This study concerns the development of a short version of a well-known and much used clinimetric instrument called the Sickness Impact Profile (SIP). The SIP is a generic measure of functional status. Based on findings of a principal components analysis of over 800 SIPs from a multi-diagnostic population, a selection of 68 items divided over 6 dimensions was made and initially tested. As no support was found for the statistical validity of the categorical structure of the original SIP, a new structure, discovered through principal components analysis, was used as the basis for selecting items. Comparison of the scores on the selection with information provided by the original SIP showed very promising results: the 68 item selection may serve as a valid short SIP-version.

3.1. Introduction

The Sickness Impact Profile (SIP) is one of the best known general health measures. In 1986 Spitzer suggested the SIP as one of three instruments to which the research concerning the methodology of research on the quality of life and functional status measurement should be limited [2]. The SIP aims at measuring the health-related functional status by assessing the behavioral impacts of sickness. It measures physical functioning as well as emotional and social aspects of functioning. SIP is intended as an outcome measure of health care, to be used for evaluation, program planning and policy formation [3]. The authors anticipated that the SIP would be used for a variety of purposes, ranging from the assessment of the general health status of a population, to following clinical progress of an individual patient [4]. The instrument has indeed been used for these purposes in studies in the United States and in several countries in Europe. In former publications we reported data on reliability and validity of the SIP [5]. It appears that in research practice, it is considered to be a reliable and valid instrument. Hence it has been highly recommended for use [2,6,7,8,9,10], and was even advised for use as a gold standard [11].

The SIP consists of 136 statements, every item describing a possible sickness impact. The items are divided over twelve categories, which cover sleep and rest, emotional behavior, body care and movement, home management, ambulation, social interaction, mobility, alertness behavior, communication, work, recreation and pastimes, and eating. The constructors of the instrument paid great attention to the comprehensiveness of the catalogue of possible sickness impacts. During the construction, the aim was 'to obtain as comprehensive a sample of the impact of sickness on a persons' behavior as possible' [3]. The constructors of the SIP wanted to develop an instrument describing clinical phenomena, and emphasis was laid on a complete coverage of 'behavioral sickness impacts'. Statements were obtained from patients, health care professionals, individuals caring for patients, and from apparently healthy individuals [12]. The gathering, sorting and grouping of statements were aimed at collecting and retaining any statement that was judged useful, relevant, unique or discriminative [13]. Staff members selected, and when necessary, rewrote the items. Initially there were 1250 items describing possible impacts. Items judged as referring to a common type of activity were grouped together into categories, and through systematic selection, the final instrument was obtained [2]. This instrument provides a comprehensive clinical description of the functional status, but, due to its clinical basis, has more or less the character of a clinical check list. The fact that the SIP was originally meant to provide a descriptive profile of the functional status of an individual illustrates this. The (clinical) relevance of the items and the categorical structure were defined by staff members. The categories therefore, represent twelve clinically distinguishable aspects of functional status. For research purposes however, functional status is mostly viewed from a more global, abstract or theoretical angle. Although the internal structure of the SIP has high clinical- or face-validity and the external

construct validity of the instrument is generally accepted, the theoretical or empirical validity and relevance of the categorical structure (the internal construct validity) has never, to our knowledge, been evaluated and is still open to question. As the information provided by the SIP, is often interpreted and analyzed for research purposes by means of statistical analysis of the total score as well as scores on one or both dimensions or category scores, the empirical validity of the internal structure is a vital topic. The empirical structure could be different from the a priori categorical structure that is based on face validity. Moreover, adding valid descriptive categories of a profile, does not necessarily produce a valid measure of the total concept of functional status.

A repeated point of criticism concerning the SIP is its relatively large number of items [eg.14,15,16]. Deyo found that respondents experienced items not applying to their situation as redundant [14]. This is one of the few (slightly) negative reactions elicited by the instrument. Moreover, as far as its length is concerned, the SIP compares unfavorably to instruments measuring related concepts. On the other hand, it appears that, due to its length, the SIP provides a comprehensive description whose components (categories) are related to the way clinicians define aspects of functional status. Hence the instrument is much appreciated in clinical use. However, despite the fact that the information provided by the SIP is judged as a valid assessment of functional status, the length of the instrument appears to be a barrier in using it for research purposes. As a result the question arises whether it is possible to develop a short version of the instrument (a 'core-SIP') for research purposes so that it would yield an assessment of 'health related functional status' that is as reliable and valid as the original SIP. Evaluation of the categorical structure as suggested above, might at the same time, disclose possibilities to reduce the number of items in the SIP, without reduction of its validity as a measure of the 'health related functional status'.

The main questions to be answered in this paper therefore are:
- is the a priori categorical structure of the SIP present when tested by means of statistical methods?
- do the results of this analysis disclose possibilities to develop a shorter instrument?

To investigate these questions, principal components analysis will be used on a large multi diagnostic database to test the categorical structure. The outcome of this analysis will be used to develop a core-SIP. As the information gathered using the original SIP is broadly accepted as a valid and reliable assessment of functional status, an acceptable (short) alternative should provide information closely related to the original. Therefore a comparative analysis on the short and the original instrument will be the final step in this study.

3.2. Data and methods

3.2.1. Data

The data used in this study was obtained from studies in 10 different diagnostic groups, with a total of 2527 respondents. In all studies, the Dutch translation of the original SIP [8] was administered. Also, in all studies information was gathered on age, sex, education and duration of sickness. Table 3.1 presents the main characteristics of the populations.

To prevent bias by the unbalanced representation of diagnoses in the research population, samples were taken from the diagnostic groups. We tried to achieve a balanced cross section of the diagnoses: from every diagnostic group larger then 100 respondents, a random sample of 100 was drawn, and smaller subpopulations were totally incorporated. This resulted in a balanced, 835 respondents sample, that will be used in further analyses. As can be read from table 3.1, the mean age in this study population was 47 years, and sexes were well balanced (51% male). The mean total SIP-score was 10.8 with a standard deviation of 9.2.

The category 'work' was removed from the list. This category did not apply to many of the respondents within the population because they did not work before their sickness occurred (housewives, retired persons, students). Moreover, due to differences in the social security system between the Netherlands and the USA,

Table 3.1. Characteristics of the population used in this study

diagnoses	N*	% male	age(sd.)	136-SIP(sd.)	references
rheumatoid arthritis	100 (377)	26.0	58.2 (12)	13.4 (9.6)	17,18
ankylosing spondilitis	100 (264)	71.0	40.9 (11)	9.9 (8.0)	18,19,20
spinal cord injury	41 (41)	81.0	40.1 (21)	19.9 (8.1)	18
stroke	53 (53)	45.3	66.7 (12)	15.5 (10.8)	21
cancer	100 (103)	47.0	59.8 (13)	11.0 (9.4)	22
neuromusc. disease	100 (1104)	49.0	45.6 (15)	12.5 (9.0)	23
back/neck pain	100 (338)	55.0	41.0 (11)	4.8 (3.9)	24,25
head injury	100 (106)	73.0	27.6 (7)	10.5 (9.7)	26
hemodialysis	99 (99)	55.6	53.5 (16)	11.1 (9.3)	27
m.Crohn	42 (42)	31.0	35.2 (7)	5.9 (6.7)	19
Total	835 (2527)	51.0	47.2 (17)	10.8 (9.2)	

* number of respondents used in this study, total number of respondents available is shown between brackets

items in this category did not apply to the Dutch situation and therefore do not supply relevant information. For these reasons, in a number of studies from which data was used, 'work' was deleted beforehand, causing a large number of missing values in the analysis of the database in which all bases were combined. These three reasons lead to the decision not to include this category in the analyses.

3.2.2. Methods

To evaluate the categorical structure of the SIP, we used Principal Components Analysis (PCA). This is a multivariate technique aimed at detecting the main independent dimensions of variation in group characteristics [28]. It provides a number of components called factors that, after rotation (Varimax), give a concise description of coherent, relatively independent subgroups of variables within a dataset. Thus PCA-findings can validate the (hypothetical) structure of an instrument [29]. Interpretation of the content of the factors after rotation (Varimax) reveals the aspects to be distinguished within the information provided by the SIP, and thus within 'health related functional status'. The height of the factor loadings indicates the relevance of the items for a specific factor. Thus, by selecting the highest loading items, a selection of the most relevant items from the factorial structure will be obtained. Items with low factor loadings (less than 0.40) are judged less relevant. A factor loading of less than0 .40, indicates that an item has less than 16% of its variation involved in that factor [28].
Items with very skewed response patterns only have relevance for a very small number of respondents, or for almost all respondents and therefore hardly provide differentiation within the population. As the aim of a research- instrument is to distinguish groups within a population, extremely skewed items, from a research point of view, can be considered less relevant. As our population had a multidiagnostic character and the SIP is a generic instrument, items applying to no more than 10% or over 90% of any diagnostic subpopulation were considered skewed, and were removed.
The selection resulting after removal of skewed and low loading items, was tested for its structural robustness, by comparing factor solutions in different subpopulations, using Cattell's salient similarity index s, for comparison among factor solutions, as described by Tabachnick [29, p.642-644]. Next, the selection was compared with the original SIP, by comparing the scores obtained on both instruments. Finally, the total score on the selection was used in a regression analysis to predict the total score on the original SIP.

In the scoring procedure of the original instrument, nominal weights were attached to every item. To assess the possible loss or bias of information caused by selecting items without accounting for the weights, the influence of the weights on the SIP category- dimension and total scores was studied. Nominal weighing like it is used in the SIP, has its greatest effect when the weights differ considerably, when there is little inter-correlation among the components, and when there are a small number of components or items [30,31]. The question is whether weighing the items in

the SIP has an effect on the total score, because there is a large number of items, relatively little difference between weights, and high internal consistency. Therefore, an alternative scoring procedure was used. In this procedure the score was obtained by just adding the number of items checked in a category, dimension or the total list. Both weighted and 0/1 scores were calculated for the original SIP. The scores were correlated for every category, both physical- and psychosocial dimension and for the total instrument. The correlation coefficients found ranged from 0.94 to 0.99 (median 0.98). It was concluded that no bias would result from using a scoring procedure without weights. Therefore, the item weights were not taken into account in the selection.

3.3. Results

The structure hypothesized within 'health related functional status' by the authors, is represented by the 12 a priori categories of the SIP. Each category is supposed to measure a separate aspect of functional status on which health deviations might have an impact. The original categorical structure is used as a starting point. Given the twelve categories of the instrument, a rotated factor solution is expected to consist of twelve (eleven after removal of 'Work') major subgroups of items, to be interpreted according to the categories headings.
The results of the first Principal Components Analysis procedure did not support the a priori category structure of the SIP. The extracted [36] factors did not show any resemblance with the a priori structure. Also, when the number of factors to be extracted was limited to 12, the factors did not show resemblance to the a priori twelve categories. Apparently, the structure presupposed by the constructors within 'health related functional status' is not present in our data. The factors found could not be interpreted as homogenous and logical aspects of 'health related functional status'. We therefore decided to look for an alternative dimensional structure by taking the items which interfered with this structure out of the instrument. Nineteen skewed items were identified and removed from the list (see table 3A, appendix). A Principal Components Analysis was performed on the basis of the remaining 108 items. Twenty nine factors were extracted. As the successive height of the factor loadings did not show a clear bend, indicating a number of relevant factors to be found, 2.0% was chosen as the minimal level of variance to be explained by a factor to be called relevant. From this criterion, together with the analysis of the content of the factors, it was concluded that a six factor model appeared to offer an interpretable solution. These six factors are considered to represent global, and more or less universal aspects of functioning on which health status might have an impact. Thirty-five items had low loadings (<0.40 on any of the six factors), and were removed. Another PCA run was performed on the remaining 73 items. As a result of this run, five more items showed loadings <0.40, and thus they were dropped. Table 3B (appendix) presents the items dropped from the list due to low factor loadings. Part of these items was removed because they did not load significantly high (<0.30) on any factor, and thus apparently do

Table 3.2. Comment on items removed because of low factorloadings

category	comment
sleep/rest	Dropped because they do not fit in the structure (loading < .30). Items show low loadings, scattered over different factors.
emotional beh.	The items did load on one factor. Loadings <.40 as well as <.30. Possibly conceptually different from the other items loading on this factor.
bodycare & movement	Loadings were <.40 on three different factors. Items do not fit the structure and have no direct relation with autonomy or dependency.
houshold management	Items had low loadings on factors that were not related to the item content, so they did not fit the structure.
mobility	Loadings were very low (<.30) on different factors: they did not fit the structure.
social interaction	Items form a seperate dimension ('loss of interest & loneliness') within the factor they load on. Loadings, however, were too low to keep them in the selection.
ambulation	Loadings were low (<.40) to very low (<.30) on factors that had no logical relation to the item content.
alertness & intelligence	Items did load on a factor congruent with item content, but their factor loadings <.40.
communication	Loadings <.40. The item content is covered by items remaining in the selection.
recreation & pastimes	Items were dropped due to loadings <.40 and <.30 on different factors. They obviously did not fit the structure.
eating	Dropped because of low (.39) and very low (<.30) loading

not describe an aspect of functional status related to one of the six factors. The relatively strict criterion of 0.40, however, also lead to the removal of items with loadings just below 0.40. These items in fact do load significantly on one of the factors, and thus do contribute to the description of functional status, but they were removed just because of the high criterion chosen.
In table 3.2, for every category from which items were dropped, a global comment is presented, hypothesizing why the items loaded low on any factor.

At this point it appeared that in a subsequent PCA-run, aspects like communication and eating would disappear (as a result of low loadings). This would mean a drastic reduction of the content-validity of the selection in favor of the efficiency of a shorter list. This trade-off between validity and efficiency was decided in favor of the validity. Therefore, it was decided to stop the selection at 68 items, and maintain the broad scope on functional status that is typical for the SIP. The appendix presents the resulting selection consisting of 68 items (table 3C). factor loadings are printed behind every item. Below the content of the remaining 6 factor item selection will be described. The first factor, somatic autonomy (eigen value 11.1),

describes the level to which an individual is autonomous in his or her basic somatic functioning. Functions referred to are: getting dressed, standing, walking, eating and the fact that help is needed in those functions. Therefore, a higher score on this factor means a higher level of dependency and thus less autonomy, in this area. As heading for the second factor, mobility control (eigen value 5.0) was chosen. In this factor behavior is described that is related to the level to which an individual has control over his body. Six out of the twelve items in this factor are directly related to walking (eg.'I go up and down stairs more slowly, for example one step at a time, stop often'). The other items have to do with hand- and arm-control (eg.'I have difficulty doing handwork, for example turning faucets, using kitchen gadgets, sewing, carpentry'). A high score on this factor corresponds with a relatively low level of control over body movements. Psychologic autonomy and communication (eigen value 3.4) is the third factor which is distinguished. Here behavior is described that is associated with the level to which an individual is able to function without help of others in mental areas of functioning, including the possible sickness impacts on a persons (verbal) communication. When someone obtains a high score for this dimension, he will need help from others as far as his mental functioning and communication is concerned. The possible consequences of a health deviation on a person's functioning in relation to other persons (spouse, children, 'other people' in general) is described in the factor called social behavior (eigen value 2.3).
Sexual activity, visiting friends and activities in groups of people, among others, are mentioned in this factor. The fact that items concerning the amount of drinking and eating also loaded on this factor and not on one of the locomotor factors, indicates that these activities are influenced by their social context. Someone obtaining a high score on this factor will have problems in his or her social functioning. The next factor is called emotional stability (eigen value 2.0). This factor assesses the effect health status has on the emotional status of a respondent. Items in this factor are concerned with irritability, and/or acting disagreeable with yourself or others. A high score points at a lack of stability in, or control over the emotional status. The final factor is mobility range (eigen value 2.7): this dimension is concerned with the influence of health status on a number of usual tasks like shopping, house cleaning and taking care of personal business affairs. In this factor it is not the level of control over motor functions that is described, but the range of actions to which a person has (limited) disposition, given his or her motor control.

To test the robustness of the factorial structure, a principal components analysis was performed in different diagnostic subgroups and in that part of the total population that was not used to construct the selection. The total population was divided into four groups: locomotor complaints (rheumatoid arthritis, spinal cord lesion, back and neck-complaints, ankylosing spondylitis), cerebral problems (stroke, head injury), internal diseases (cancer, hemodialysis, Crohn's disease), and neuro-muscular diseases. Finally a PCA was performed on all respondents who were not incorporated in the sample that was selected at the start of this study (see 'data').

To compare the findings in these (sub-)populations with the structure found in the multidiagnostic study population, the level of agreement between factors found in the study population, and the factor solution found in the other (sub-)populations was assessed. An indicator of similarity between factor solutions is Cattell's salient similarity index s. Cattell also provided estimates of probability of similarity (p): if a value of s exceeds that of vs, then the factors compared are reliably similar [29]. In table 3D (appendix) values of s, vs and p are presented. All s-indexes found were significantly higher than the vs's found, and probabilities of similarity all indicate significance ($p < 0.05$). Thus, the factor solution as it is found within the study population, is similarly found within all four subpopulations. This indicates that the six factor structure is robust over different diagnostic groups, thus supporting the generic character of the selection.
After the selection had been completed, the new instrument was tested for its reliability by calculating Cronbach's α for every factor and for the total selection. Sufficiently high alpha's were expected because the technique used (principal components analysis) aims at distinguishing homogenous subgroups of items.

Generally an α of 0.70 or more is considered acceptable. As can be seen from table 3.3, in every dimension found within the selection and in the total list, this criterion was met. The α 's of the selection are slightly better than internal consistency coefficients of the original SIP, found in literature [4]. To compare the information provided by the original SIP to information gathered by the selection, total scores and dimension scores on the SIP were related to the total scores and the scores on the factors of the selection using regression analysis. Regression formulas were calculated to predict the total SIP score from the total score on the selection. Next the original SIP total score was predicted from the scores on the six new factors, and finally, the original dimension scores were predicted from the six new factor scores. The relation between the two total scores is reflected in the regression formula: $SIP136 = 0.35 + (.97 \times SIP68)$ ($R^2 = .94$).

Table 3.3. Internal consistency of the 68-item-selection

α = Cronbach's α; n = number of items

	α	n
Somatic autonomy	.78	17
Mobility control	.85	12
Psych. aut. & communic.	.77	11
Mobility range	.79	10
Social behavior	.81	12
Emotional stability	.72	6

Table 3.4. Beta coefficients of scores on the six factors within the short list in a regression analysis of the scores on the factors on the SIP total score and SIP dimension scores.

(n=835)

Independent	Dependent		
	total score SIP	phys. dim. score	psy.soc. dim. score
SA	.31	.62	.04
MC	.23	.39	---
PAC	.22	-.03	.55
SB	.24	.06	.17
ES	.17	---	.42
MR	.28	.18	.12
constant [1]	.75	.14	.24
R^2 [2]	.96	.96	.88

[1] constant in the regression equation with the beta coefficients in this column.
[2] R^2 of the regression equation, representing the 'fit' of the equation.

Table 3.4 presents the results of the other analyses. From the R^2's it can be concluded that the fit for all three regression equations is good, so only a limited amount of information from the SIP total and dimension scores is lost during the item selection procedure. The total score and scores on the factors of the selection were used to predict the original SIP total score and the dimension scores. Beta's range from 0.17 to 0.31 for all six factors in predicting the SIP total score. The prediction of the SIP physical dimension score is mainly based on scores on Somatic Autonomy, Mobility Control and Mobility Range. The Psychosocial dimension score is mainly predicted from scores on Psychic Autonomy and Communication, Emotional Stability and Social Behavior. The same regression analyses were performed using that part of the total population that was not used to develop the short list (see table 3.1). Findings in this population were to a very high degree identical to the findings in the sample used to select items.
The relatively high correlations between the six factors (range: 0.18 to 0.54, mean: 0.33) and the findings presented in table 3.4 suggest a two dimensional structure in the selection that is congruent with the dimensional structure in the original SIP: a physical dimension and a psychosocial dimension. This finding warranted the search for second order factors within the selection. So a Principal Components Analysis was performed on the score on the six factors of the short list. To compare the internal structure of the selection with the dimensional structure of the original instrument, a PCA was also performed on the scores on the original SIP categories. Results are presented in table 3.5.

TABLE 3.5. Second order factors (Varimax rotated) of original SIP category scores (except 'work') and the scores on the six factors in the selection.

Loadings >.40 are in bold print

	Factor 1	Factor 2
Original SIP		
Sleep/Rest	**.60**	.34
Emotional behavior	.17	**.70**
Bodycare & movement	**.80**	.15
Household management	**.78**	.24
Mobility	**.81**	.22
Socail interaction	.38	**.66**
Ambulation	**.79**	.11
Alertness & intelligence	.05	**.80**
Communication	.25	**.56**
Recreation & pastimes	**.57**	**.44**
Eating	**.55**	.17
68 Selection		
Somatic autonomy	**.78**	.01
Mobility control	**.65**	.30
Psych. aut. & comm.	.18	**.75**
Social behaviour	**.58**	**.48**
Emotional stability	.06	**.81**
Mobility range	**.85**	.11

These findings confirm the above suggestion of the dimensionality of the selection. In both instruments two factors were found. Also in both lists, the two factors could be interpreted as a physical dimension on the one hand and a psychosocial dimension on the other hand. Moreover, in both factor solutions one category or factor appears to be related both to the physical and to the psychosocial factor. In the original SIP this is the category 'Recreation and pastimes'. In the selection, it is 'Social Behavior' that takes this intermediate position. Apparently within 'health related functional status' as measured by the original SIP and the selection, three main dimensions can be distinguished: a physical, a psychosocial and an intermediate dimension that is related to the physical, as well as to the psychological aspects of functioning. This intermediate category is concerned with social and recreational activities. These three dimensions show congruence with the three dimensions of health distinguished by the WHO health definition (physical, psychological and social health).

3.4. Discussion

The Sickness Impact Profile (SIP) is known to provide valid and reliable information on the functional status of respondents. In the introduction of this paper it was argued that different criteria are to be considered when an instrument is meant for clinical use than when it is meant for research purposes. During the construction, emphasis lay on the content validity and the face validity of the instrument. This lead to a comprehensive and clinically relevant profile of functional status. The validity and relevance of the aspects distinguished by the twelve SIP categories for research purposes was the first topic in this study. Contiguous, as the SIP is regarded as too long by several researchers, the possibilities to develop a short 'core-SIP' for research purposes were explored. Two main questions are addressed in this study: (a) is the a priori categorical structure of the SIP present when tested by means of statistical methods, and (b) does the structure within the information provided by the SIP disclose possibilities to select a core instrument providing approximately the same information as the original instrument does? To answer the first question principal components analysis was used in a multidiagnostic population. No statistical support was found for the validity of the twelve factor structure of the original instrument. We assume that the use of unweighted items in this procedure, would not cause much bias, as weighted and unweighted scores on categories, dimensions or the total SIP showed extremely high correlations with scores obtained using weighted items. With regard to the second question, promising results were obtained. Based on findings of a principal components analysis after removal of skewed items, a selection of 68 SIP items was found. The information provided by this selection appears closely related to the information provided by the SIP. The fact that the selection procedure was performed using unweighted items might have caused a bias in the selection of items.

The high correlations between (unweighted) selection scores and weighted SIP scores, however, suggests that dropping the item weights did not cause bias. A study of the content of the items removed because of a skewed answering pattern, revealed that there were two main explanations for this skewness. The first explanation was that the skewed items describe very severe states of illness, eg. conditions of being bedridden ('I stay lying down most of the time', 'I stay within one room', 'I am eating no food at all, nutrition is taken through tubes or intravenous fluids') or conditions in which the possibilities to communicate are severely hindered (eg. 'I communicate mostly by gestures, for example, moving head, pointing, sign language', or 'I sit around half-asleep'). In most populations used in this study, very ill people were not included because they were not able to participate and to complete a long questionnaire like the SIP. This under-representation of severely ill respondents in our population might have caused a bias in favor of items describing not extremely severe sickness impacts. The selection therefore, might not be able to distinguish between seriously ill and very seriously ill respondents. However, in studies using instruments like the SIP, very seriously ill people will most probably not participate because they are not able to. Therefore, the loss

of these items will not affect the applicability of the selection to a great extend. A second factor that might have caused the skewness of the items, is the fact that a number of these items describe behavior that is not considered to be socially acceptable (eg. 'I refuse contact with family members, for example turn away from them.' and 'I isolate myself as much as I can from the rest of the family'.) Respondents will hesitate to admit that items like these describe their own behavior. Therefore, only very few respondents will have checked those items, causing skewness. As this aspect of social desirability will also bias the scores on the original SIP, the loss of these items will also not detract from the information provided by the selection compared to that provided by the SIP. Forty items were removed because they loaded <.40 on any factor in the principal components analysis. Low loadings were mainly caused by the fact that items did not fit into the structure defined by the six factors found in the selection. The fact that a minimal loading of .40 was chosen as a criterion resulted in the removal of a number of items on arbitrary grounds. Items loading less than .40 on any factor were dropped in spite of the fact that they might fit logically into the new factorial structure. The high correlation (.97) between the total scores of both lists however, suggests that dropping these items did not cause the loss of information vital to the total concept of health related functional status as measured by the SIP. Findings of a regression analysis indicate that the original SIP-total score can be almost exactly predicted from the information provided by the selection total score. Within the frame of this study this finding in a regression analysis was to be expected, as the selection was drawn from the original SIP, and a high correlation exists between the two total scores. Therefore, further study, using the selection as a separate instrument, and the original instrument in the same situation, is needed to reveal more about the empirical relation between the selection and the SIP. Moreover, before the selection can be taken as a valid international alternative for the SIP, the analyses performed in this study should be repeated using valid versions of the SIP in other languages (eg. English, Danish, Swedish, Spanish, French, German). The (categorical or factor)structure of both lists shows a remarkable parallel: within both lists a physical and a psychological dimension can be distinguished next to a 'social' dimension that loads approximately equal on the two former factors. The internal consistency of the short list (Cronbach's α of the factors and of the total list) appears to be better than the internal consistency of the categories in the original SIP. The relatively high α's in the dimensions of the selection indicate that the technique used to construct these dimensions (principal components analysis) succeeded in distinguishing homogenous groups of variables. Moreover, as Cronbach's α is known to be positively biased by a large number of items, the higher alpha's found in the selection (that consists of half as many items as the original instrument), constitute an emphatic support of the reliability of the selection. A preliminary conclusion based on statistical findings in this study is that in this population, the selection is a good alternative for the original list.

An important point in the discussion is the generalizability of the findings. In other words: is the selection only an alternative for the original instrument in our population or can it serve as a general generic alternative for the SIP? The population used in

this study mainly consisted of chronic and lengthy conditions, including pain or locomotor disorders, neurological, internal and cardiovascular disorders. In previous research we found these to be the most important areas in which the SIP was used [4]. From this it can be derived, that the findings in this study might apply to most situations in which the SIP is used. As mentioned before, the removal of skewed items might have harmed the possibilities of the list to distinguish between severely ill and very severely ill respondents. However, as very severely ill people will not be able to fill out a lengthy questionnaire like the SIP, the applicability of the selection, as far as level of illness in the population is concerned, is not worse for the selection than it is for the original instrument.
A crucial condition in the discussion about the validity and reliability of the selection as an alternative for the SIP is that all data on the selection are based on information gathered using the original SIP. This implies that the context of the total SIP might have influenced the findings on the selection in this study. Using the selection as an instrument without the context of the total SIP will have to prove the quality of the selection as a measure of health related functional status. Comparison of separate findings from the SIP and from the selection, is needed to provide more conclusive information on the validity and reliability of the selection as an alternative for the SIP. The conclusion, based on the findings in this study, however, is that the 68 item selection provides a very promising short generic alternative for the original Sickness Impact Profile.

Appendix

Table 3A. Skewed items

Category	
SR	- I am sleeping or dozing most of the time - day and night.
	- I sit around half-asleep.
EB	- I have attempted suicide.
BcM	- I stay lying down most of the time.
	- I spend most of the time partly undressed or in pajamas.
MOB	- I stay within one room.
	- I am staying in bed most of the time.
SI	- I make many demands, for example, insist that people do things for me, tell them how to do things.
	- I isolate myself as much as I can from the rest of the family.
	- I am paying less attention to the children.
	- I refuse contact with family members, for example turn away from them.
AMB	- I walk up or down stairs only with assistance from someone else.
Com	- I communicate mostly by gestures, for example, moving head, pointing, sign language.
	- My speech is understood only by a few people who know me well.
	- I carry on a conversation only when very close to the other person or looking at him.
E	- I eat no food at all but am taking fluids.
	- I just pick or nibble at my food.
	- I do not feed myself at all, but must be fed.
	- I am eating no food at all, nutrition is taken through tubes or intravenous fluids.

Table 3B. Items dropped because of low factorloadings

S/R	- I sit during much of the day. - I spend much of the day lying down in order to rest. - I sleep or nap more during the day. - I lie down more often during the day in order to rest. - I sleep less at night, for example, wake up too early, don't fall asleep for a long time, awaken frequently.
EB	- I say how bad or useless I am, for example, that I am a burden on others. - I get sudden frights. - I laugh or cry suddenly. - I talk about the future in a hopeless way. - I keep rubbing or holding areas of my body that hurt or are uncomfortable. - I act nervous or restless. - I often moan and groan in pain or discomfort.
BcM	- I have trouble getting shoes, socks or stockings on. - I change position frequently. - I hold on to something to move myself around in bed.
HH	- I do work around the house only for short periods of time or rest often. - I am not doing any of the maintenance or repair work that I would usually do in my home or yard.
Mob	- I am not now using public transportation. - I am staying in bed more. - I am only going to places with restrooms nearby.
SI	- I am going out less to visit people. - I am avoiding social visits from others. - I show less affection. - I show less interest in other people's problems, for example, don't listen when they tell me about their problems, don't offer to help. - I often express concern over what might be happening to my health. - I stay alone much of the time. - I am not doing the things that I usually do to take care of my children or family.
Amb	- I walk only with help from someone. - I get around only by using a walker, crutches, cane, walls or furniture.
AI	- I have more minor accidents, for example drop things, trip and fall, bump into things. - I sometimes behave as if I were confused or disoriented in place or time, for example, where I am, who is around, directions, what day it is.
Com	- I don't write except to sign my name. - I am understood with difficulty. - I often lose control of my voice when I talk, for example, my voice gets louder or softer, trembles, changes unexpectedly.
RP	- I am not doing any of my usual inactive recreation and pastimes, for example watching TV, playing cards, reading. - I am not doing any of my usual physical recreation or activities. - I am cutting down on some of my usual physical recreation or activities. - I am doing more inactive pastimes in place of my other usual activities.
E	- I am eating special or different food, for example, soft food, bland diet, low-salt, low-sugar. - I feed myself but only by using specially prepared food or utensils.

Table 3C. 68 item selection

Factorloadings between brackets

Somatic autonomy

1. I get around in a wheelchair.(.73)
2. I get dressed only with someone's help.(.70)
3. I do not move into or out of bed by myself, but am moved by a person or mechanical aid.(.67)
4. I stand up only with someone's help.(.65)
5. I do not fasten my clothing, for example require assistance with buttons, zippers, shoelaces.(.63)
6. I do not walk at all.(.57)
7. I do not use stairs at all.(.57)
8. I make difficult moves with help, for example, getting into or out of cars, bathtubs.(.53)
9. I do not bathe myself completely, for example, require assistance with bathing.(.49)
10. I do not bathe myself at all, but am bathed by someone else.(.48)
11. I do not have control of my bladder.(.47)
12. I am very clumsy in body movements.(.45)
13. I do not have control of my bowels.(.43)
14. I feed myself with help from someone else.(.43)
15. I do not maintain balance.(.42)
16. I use bedpan with assistance.(.41)
17. I change position frequently.(.40)

Mobility control

1. I go up and down stairs more slowly, for example, one step at a time, stop often.(.68)
2. I walk shorter distances or stop to rest often.(.63)
3. I walk more slowly.(.60)
4. I use stairs only with mechanical support, for example , handrail, cane crutches.(.59)
5. I walk by myself but with some difficulty, for example, limp, wobble, stumble, have stiff leg.(.57)
6. I kneel, stoop or bend down only by holding on to something.(.56)
7. I do not walk up or down hills.(.49)
8. I get in and out of bed or chairs by grasping something for support, or using a cane or walker.(.47)
9. I stand only for short periods of time.(.45)
10. I dress myself, but do so very slowly.(.44)
11. I have difficulty doing handwork, for example turning faucets, using kitchen gadgets, sewing, carpentry.(.44)
12. I move my hands or fingers with some limitation or difficulty.(.44)

Psychologic autonomy and communication

1. I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things.(.67)
2. I have difficulty doing activities involving concentration and thinking.(.65)
3. I react slowly to things that are said or done.(.61)
4. I make more mistakes than usual.(.58)
5. I do not keep my attention on any activity for long.(.57)
6. I forget a lot, for example, things that happened recently, where I put things, appointments.(.56)
7. I am confused and start several actions at a time.(.49)
8. I do not speak clearly when I am under stress.(.47)
9. I have difficulty speaking, for example, get stuck, stutter, stammer, slur my words.(.46)
10. I do not finish things I start.(.43)
11. I am having trouble writing or typing.(.41)

Social behavior

1. My sexual activity is decreased.(.57)
2. I am cutting down the length of visits with friends.(.51)
3. I am drinking less fluids.(.51)
4. I am doing fewer community activities.(.51)
5. I am doing fewer social activities with groups of people.(.49)
6. I am going out for entertainment less often.(.49)
7. I stay away from home only for brief periods of time.(.47)
8. I a eating much less than usual.(.45)
9. I a not doing heavy work around the house.(.44)
10. I do my hobbies and recreation for shorter periods of time.(.43)
11. I am doing less of the regular daily work around the house than I would usually do. (.43)
12. I am cutting down on some of my usual inactive recreation and pastime, for example, watching TV, playing cards, reading.(.42)

Emotional stability

1. I often act irritable toward those around me, for example, snap at people, give sharp answers, criticize easily.(.66)
2. I act disagreeable to family members, for example, I act spiteful, I am stubborn.(.62)
3. I have frequent outbursts of anger at family members, for example, strike at them, scream, throw things at them.(.57)
4. I act irritable and impatient with myself, for example, talk badly abut myself, swear at myself, blame myself for things that happen.(.44)
5. I am not joking with family members as I usually do.(.42)
6. I talk less with those around me.(.42)

Mobility range

1. I am not doing any of the shopping that I would usually do.(.65)
2. I am not going into town.(.64)
3. I am not doing any of the house cleaning that I would usually do.(.58)
4. I am not doing any of the regular work around the house that I would usually do.(.57)
5. I stay home most of the time.(.55)
6. I am not doing any of the clothes washing that I would usually do.(.53)
7. I am not going out to visit people at all.(.52)
8. I am getting around only within one building.(.47)
9. I have given up taking care of personal or household business affairs, for example paying bills, banking working on budget.(.45)
10. I do not get around in the dark or in unlit places without someone's help.(.44)

Table 3D. Comparison among solutions, probabilities (p) for s > vs

diagnoses		som aut	mob cont	psych.aut & comm	social behav	emot stab	mob range	mean s
locomotor	s	**.83**	**.80**	**.78**	**.50**	**.80**	**.89**	**.77**
complaints	vs	.63	.76	.76	.34	.76	.76	
	p	.00	.00	.00	.05	.00	.00	
cerebral	s	**.74**	**.67**	**.78**	**.40**	**.67**	**.89**	**.79**
problems	vs	.63	.51	.76	.34	.51	.76	
	p	.00	.02	.00	.05	.02	.00	
internal disease	s	**.58**	**.67**	**.90**	**.96**	**.91**	**.95**	**.69**
	vs	.51	.51	.76	.76	.76	.76	
	p	.00	.02	.00	.00	.00	.00	
neuro-muscular	s	**.74**	**.80**	**.84**	**.67**	**1.0**	**.67**	**.82**
disease	vs	.63	.76	.76	.51	.76	.51	
	p	.00	.00	.00	.02	.00	.02	
mean s		**.72**	**.73**	**.82**	**.63**	**.84**	**.85**	
Non sample	s	**.71**	**.74**	**.90**	**.78**	**.67**	**.56**	**.73**
population	vs	.63	.51	.76	.76	.51	.51	
(n=1551)	p	.00	.02	.00	.02	.02	.02	

s: Cattell's salient similarity index s
vs: criterion for similarity, if s exceeds vs, then the factors are reliably similar.
p: esimate of probability values for s
source: Tabachnick BG., Fidell LS., Using multivariate statistics, second edition, Harper & Row New York 1989, p.642-644.

Literature

1. Feinstein AR. Clinimetrics. New Haven: Yale University Press, 1987.
2. Spitzer WO. Keynote address: State of science 1986: quality of life and functional status as target variables for research. J. Chron. Dis. 1987, 40(2): 465-474.
3. Bergner M, RA Bobbit, WB Carter, BS Gilson. The Sickness Impact Profile: development and final revision of a health status measure. Medical Care 1981, 8: 787-805.
4. Gilson BS, M Bergner, RA Bobbit, WB Carter. The sickness impact profile: final development and testing 1975-1978. Seattle, Washington: University of Washington, Department of Health Services, 1979.
5. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks. Sickness Impact Profile: the state of the art of a generic functional status measure, Soc.Sci.Med. 1992, 8: 1003-1014.
6. Deyo RA. Measuring functional outcomes in therapeutical trials for chronic disease. Cont. Clin Trials 1984, 5: 223-240.
7. Patrick DL. A cross-cultural comparison of health status values. Am.J. of Public Health 1985, 75(12): 1402-1407.
8. Luttik A, H Jacobs, LP de Witte. Een nederlandse versie van de Sickness Impact Profile (A Dutch version of the Sickness Impact profile). Vakgroep Huisartsgeneeskunde, Rijksuniversiteit Utrecht, 1985.
9. Sullivan M, M Ahlmén, B Archenholtz, G Svensson. Measuring health in rheumatic disorders by means of a Swedish version of the SIP. Results from a population study. Scand. J. Rheumatology 1986, 15(2): 193-200.
10. Bowling A. Measuring Health: a review of quality of life measurement scales. Philadelphia: Open University Press, 1991.
11. McDowell I, C Newell. Measuring Health: A guide to rating scales and questionnaires. New York: Oxford University Press, 1987.
12. Gilson BS, JS Gilson, M Bergner, RA Bobbit, S Kressel, WE Pollard, M Vesselago. The Sickness Impact Profile, development of an outcome measure of health care. A.J.P.H. 1975, 65(12): 1304-1310.
13. Bergner M, RA Bobbit, WE Pollard, DP Martin, BS Gilson. The Sickness Impact Profile, Validation of a health status measure. Medical Care 1976, 1: 57-68.
14. Deyo RA, TS Inui, JD Leininger, SS Overman. Measuring functional outcomes in chronic disease: A comparison of traditional scale and a self-administered health status questionnaire in patients with rheumatoid arthritis. Medical Care 1983, 21: 180-192.
15. Essink-Bot ML, MPMH Rutten-van Mölken. Het meten van de gezondheids-toestand (Measuring the state of health). Rotterdam: Erasmus Universiteit, Instituut Maatschappelijke Gezondheidszorg, 1991.
16. Fletcher AE, EJ Dickinson, I Philp. Review: audit measures: quality of life instruments for everyday use with elderly patients. Age and Ageing 1992, 21: 142-150.
17. de Witte LP, DJP Tilli, AJG Ticheler, BAC Winants, FG van der Horst, Sj van der Linden. Leven met een reumatische aandoening, een onderzoek naar de kwaliteit van leven bij 372 mensen met een reumatische aandoening (Living with a rheumatic disorder: research on the quality of ilfe of 372 people with a rhematic disorder). Hoensbroek: IRV, 1989.
18. de Witte LP. After the rehabilitation center: a study into the course of functioning after discharge from rehabilitation. Amsterdam: Swetz & Zeitlinger, 1991.
19. Janssen M. Personal networks of chronic patients. PhD. thesis, Maastricht: University of Limburg, 1992.
20. Hidding A, Sj van der Linden, M Boers, X Gielen, A Kester, LP de Witte, B Dijkmans, J Moonenburgh. Is group physical therapy superior to individual therapy in ankylosing spondylitis. A randomized controlled trial. Arthritis Care and Research, submitted.
21. Visser-Meily A, MGPM Geerts. Beloop en beperkingen van CVA-patiënten gedurende het eerste jaar na het CVA (The course and limitations of functioning of cerebrovasculair accident patients during the first year after the cerebrovascular accident). presentation at the 'fall symposium VRA'. Hoensbroek: 30-10-1992.

22. Courtens AM. Kenmerken van zorg en kwaliteit van leven bij patiënten met kanker (Characteristics of Care and Quality of Life in cancer patients), PhD. thesis, Maastricht: University of Limburg, 1993.
23. Terpstra Sj, LP de Witte, FG van der Horst. De rol van een patiënten organisatie in de zorg voor chronisch zieken en hun gezin (The Role of a Patient Organisation in Care for the Chronically ill and their Families). Hoensbroek: IRV, 1991.
24. Keijzers JFEM. The efficacy of backschools: empirical evidence and its impact on health care practice. PhD. thesis, Maastricht: University of Limburg 1991.
25. Koes BW. Efficacy of manual therapy and physiotherapy for back and neck complaints. PhD. thesis, Maastricht: University of Limburg, 1992.
26. Peeters JME, LP de Witte, JCM van Haastregt. Het dagelijks funktioneren en revalidatiebehandeling van mensen met een traumatisch hersenletsel (Daily Functioning and Rehabilitation Treatment of people with a traumatic Brain Injury). Hoensbroek: IRV, 1992.
27. Severens J. Effectiviteit van EPO; kwaliteit van leven van hemodialysepatiënten met en zonder erytropoëtine (Effectiveness of EPO: Quality of Life of heamodyalysis patients with and without erythropoietin). masters thesis, Maastricht: University of Limburg, 1989.
28. Rummel RJ. Applied Factor Analysis. Evanston: Northwestern University Press; 1970.
29. Tabachnick BG, LS Fidell. Using multivariate statistics. New York: Harper & Row, Publishers, Inc.; 1989.
30. Ghiselli EE. Measurement theory for the behavioral sciences. San Francisco: WH. Freeman and Company 1981.
31. Guilford JP. Fundamental statistics in psychology and education. New York: McGraw-Hill International Book Company, 1981.

Chapter 4

The Sickness Impact Profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility

A.F de Bruin[1], M. Buys[1], L.P. de Witte[1,2], J.P.M. Diederiks[1,2].

[1] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands.
[2] IRV, Hoensbroek, The Netherlands.

This paper was published in the Journal of Clinical Epidemiology 1994, vol. 47, no. 8: 863 - 871.

Abstract

In previous research a short version of the Sickness Impact Profile (SIP136) was developed, containing 68 items. This SIP68 is intended as a short generic alternative to the original SIP. High reliability of the SIP68 was reported when it was extracted from the SIP136. This paper is a report on the first reliability testing of the SIP68 administered as an independent instrument without the context of the SIP136. To establish the test-retest reliability and the internal consistency of the new instrument, 51 patients of an outpatient department of rheumatology completed the SIP68 twice, with an interval of 48 hours.
To compare the performance of the independent SIP68 with the SIP68 extracted from the SIP136, the SIP136 also was completed two times by the same 51 respondents. Test-retest reliability for both administration types was assessed by means of the Intraclass Correlation Coefficient and the Jaccard's Similarity Ratio. Internal consistency was assessed by means of Cronbach's α. The reliability appears to be high in both the independent SIP68 as well as the extracted SIP68. Moreover, the reliability of the independent SIP68 appears to be as high as for the SIP136. These findings were very encouraging, indicating that the SIP68 may very well serve as a generic alternative to the SIP136.

4.1. Introduction

In the last two decades a wide variety of general health measures have become available for use in health services research. One of these is the Sickness Impact Profile (SIP) [1]. This instrument is well known as a generic measure of health related functional status. The SIP is designed to be broadly applicable across types and severities of illness and across demographically and culturally diverse groups. It is intended for use in measuring the outcomes of care in health surveys, in program planning, in policy formation and in monitoring patients progress [1]. The SIP is known to be a valid and reliable instrument [2,3,4,5,6] McDowell & Newell, and Bowling highly recommend it for use in clinical and survey research [7,8].
However, one of the major drawbacks of the SIP mentioned in literature is its length. Compared to related instruments the SIP is long. In previous research, therefore, a short generic alternative for the SIP was developed, containing 68 of the original SIP-items divided over 6 categories (see appendix) [9]. For reasons of clarity, 'SIP136' will be used in this paper when the original instrument is meant, and the short SIP version will be indicated as 'SIP68'.
Preliminary evaluation of the SIP68 showed encouraging results for the short version as a valid and reliable generic alternative to the SIP136 [9]. A crucial aspect of these findings, however, was that all data on the SIP68 were derived from administrations of the SIP136. A logical next step therefore, is to investigate whether the psychometric properties of the SIP68 when it is administered on its own, differ from the psychometric characteristics of the SIP68 extracted from administration of the SIP136. Therefore, this paper reports on the first empirical testing of this short generic general health measure. A study is described in which the SIP68 is administered several times to visitors of an outpatient department of rheumatology. The reliability of both administration forms of the short list (SIP68 independently, and SIP68 extracted from the SIP136) was studied and compared. First, however, the SIP68 will briefly be introduced.

4.2. SIP68

As mentioned above, the SIP68 was developed from the original Sickness Impact Profile. It was constructed by selecting items from the original instrument. The selection was based on findings of principal components analysis in a multi diagnostic database containing over 2000 completed SIP136's. In an earlier publication we elaborately reported on the methods and findings during the construction of the SIP68 [9]. This paragraph will give a brief description of the instrument.
The SIP68 contains 68 items, every item being a statement on behavior. Just like in the original instrument, respondents are asked to check those items that both apply to their situation on the day they fill out the list and are related to their health status. Thus, the instrument assesses the behavioral impact of (bad) health status. The items are divided over 6 categories: somatic autonomy (17 items), mobility control (12 items), psychic autonomy & communication (11 items), social behavior (12 items),

emotional stability (6 items) and mobility range (10 items). The SIP68-items are presented in the appendix. The first category, 'somatic autonomy' (SA), describes the level to which an individual is autonomous in his or her basic somatic functions, like: getting dressed, standing, walking, eating, and the fact that help is needed in those functions. A higher score on SA means a higher level of dependency and thus less autonomy. Items in 'mobility control' (MC) describe behavior that is related to the level of control which an individual has over his body. Items in this category are related to walking, or have to do with hand and arm control. A high score on this category corresponds with a relatively low level of control over body movements. 'Psychic autonomy and communication' (PAC) is the third category. Here behavior is described that is associated with the level to which an individual is able to operate without help of others in mental areas of functioning, including the possible sickness impacts on a person's (verbal) ability to communicate. Thus, when someone obtains a high score for this category, he will need help from others as far as his mental functioning and communication abilities are concerned. The possible consequences of a health deviation on a person's functioning in relation to other persons (spouse, children, 'other people' in general) is described in the category called 'social behavior' (SB). Sexual activity, visiting friends, and activities in groups of people, among others, are mentioned in this category. Someone who obtains a high score on this dimension will have problems in his or her social functioning. The next category is 'emotional stability' (ES). This category assesses the effect health status has on the emotional status of a respondent. The 6 items in this category are concerned with irritability, and/or acting disagreeable with yourself or others. A high score points at a lack of stability in, or a lack of control over the emotional status. The final category is 'mobility range' (MR): this category is concerned with the influence of health status on a number of usual tasks like shopping, house cleaning and taking care of personal business affairs. It is not the level of control over motor functions that is described, but the range of actions to which a person has (limited) disposition, given his or her health status. The score on the SIP68 and its separate categories is calculated by adding the total number of items checked.

4.3. Subjects and methods

4.3.1. Subjects

The data was gathered during three days at an outpatient rheumatology clinic at a general hospital in Venlo, The Netherlands. At an outpatient clinic, not all visitors actually have a (rheumatic) disease. To obtain a homogeneous group of respondents, the presence of manifest (rheumatic) health problems at the first assessment (t1), was an inclusion criterion. Therefore, respondents who did not check any SIP68 item, who assessed their own health as good to very good, and who were judged as having a good to very good functional status by the rheumatologist, were excluded from participation after t1. At t1 104 people visiting the outpatient clinic were asked to cooperate. Nineteen of these outpatients refused to participate: two because of

illiteracy, the remaining 17 for different reasons (eg. lack of time or interest). Ten respondents were excluded because they did not check any item. They judged their health as good to very good, and their functional status was judged as good to very good by the rheumatologist. Eight others had much trouble completing the instrument, and therefore were excluded from participation. The difficulties of filling out the list were caused by physical limitations or difficulty in understanding the instructions. The remaining 67 respondents were handed a second short instrument, and were asked to complete and return this 48 hours later (second assessment, t2). Fifty-one respondents agreed to cooperate. Two of these respondents did not return the list at t2. At t3 and t4 (third and fourth assessment), however, they agreed to participate.
In table 4.1 characteristics of the study group and of the group that refused to participate after t1 are presented. As can be seen in this table, there are only minor differences between the remaining 51 respondents and the 16 non-respondents as far as age, sex, self assessment of health are concerned. Also, the physician's assessment of functional status and the percentage of rheumatoid arthritis patients did not differ between respondents and non-respondents. The diagnoses that are represented in the study group are presented in table 4.2. All respondents had the Dutch nationality.

Table 4.1. Main subject characteristics

	n	age[1]	%	% ra[2]	health[3]	functional status[4]
study group	51	57	69	47	3.3	3.2
refusal group	16	55	63	43	3.1	3.2

[1] mean age
[2] percentage of reumatoid arthritis patients
[3] mean of self assessment on Likert scale (1=very bad, 5=very good)
[4] mean rheumatologist assessment of functional status (1=very bad, 5=very good)

Table 4.2. Diagnoses in the study group

Diagnoses	N	(%)
rheumatoid arthritis	24	(47%)
arthritis	5	(10%)
ankylosing spondilitis	5	(10%)
primary arthroses	4	(8%)
m. Reiter	3	(6%)
ostheoporosis	3	(6%)
diverse diagnoses	7	(14%)

Scheme 4.1. Study design

t1 - 48 hours - t2 - 14 days - t3 - 48 hours - t4

t1, t2 : SIP68 administered
t3, t4 : SIP136 administered

4.3.2. Methods

Reliability can be conceptualized and measured in a number of different ways [10,11]. In this paper, reliability assessment was confined to test-retest reliability and internal consistency. In order to assess reliability, the SIP136 and the SIP68 were both administered twice to a group of patients with a rheumatic disease, according to scheme 1.

At t1, a junior-researcher asked all outpatient-clinic visitors to cooperate. They were informed about how much time their participation would take, and they were told that whether they cooperated or not would not influence their treatment by the rheumatologist. Another junior researcher introduced the instrument and was present when respondents filled out the list in a separate room. Every respondent received a second short list and a return envelope. They were asked to fill this out 48 hours after t1, and return it. As rheumatic diseases are known to vary through the day, respondents were asked to complete the questionnaire at approximately the same time of the day as they filled it out the first time. The requested date and time were noted on the return envelope. At t3, 14 days after t2, the respondents were visited at home by one of the same two junior researchers present at t1, and the SIP136 was administered. After completing the list, another SIP136 in a return envelope was handed over, with the request to complete it at approximately the same time of the day, and return this 48 hours later (t4). Again the requested time and date were noted on the return envelope.

Through an additional short questionnaire at t1, t2, t3 and t4, subjects provided self-assessments of their level of functioning (physical, psychological, social), happiness and the severity of their disease on Likert scales. At t1, after a respondent had agreed to cooperate, data was collected on age, sex, civil status and education by means of a short list of questions. Also at t1, the rheumatologist was asked to note the diagnoses and to assess the functional status of the respondents that cooperated on a five point Likert scale.
To assess the test-retest reliability, it is important that respondents do not exactly remember the answers they gave the first time they filled out the list. Although the SIP consists of statements rather than of questions, and respondents hence would be more likely to indicate their actual state of health than try to remember the

responses they gave the last time [11], a 'washout period' between two administrations is necessary. During the development of the SIP136, test-retest reliability was assessed using a 24 hours interval [2]. As the SIP68 contains half the number of items compared to the original, the exact answers given will be more easy to remember. Therefore, to be able to compare findings in this study to test-retest data on the original instrument, an interval of 48 hours between t1 and t2, and t3 and t4 was chosen. The interval of 14 days between t2 and t3 was chosen to be sure that the answers given at t2 were forgotten and the situation at t3 and t4 would be the same as at t1 and t2 respectively. Findings at t3-t4 would then also be comparable to findings at t1-t2 and to test-retest data concerning the SIP136 from the literature. During the two weeks washout period there would also be sufficient time to receive the completed lists and to make arrangements for the visits at home.
Although rheumatic diseases may vary within the course of one day, when the disease process is observed over a number of days, it is relatively stable (at least over this period of two weeks). Hence, no substantial changes in overall functional status and thus in SIP68 scores were expected over a period of two weeks. To control for possible changes in health status, the self assessment Likert scales were used.
To assess the test-retest reliability, the Intraclass Correlation Coefficient (ICC) was computed [12] using the scores obtained at the four administrations. The ICC is an estimate of reliability to be calculated over more than two assessments. It ranges from 0 to 1, and it is derived from the information in an Anova table. The ICC is a coefficient of reliability with respect to the total instrument (or category). As The SIP68 can also be considered to provide a profile of disfunctioning when the actual disfunctions (items checked) are considered, reliability also has to be evaluated at item level.
The Jaccard's Similarity Ratio (JSR) is a measure of reliability at item level [13,14]. In this ratio the number of items checked on both occasions is divided by the sum of the number of items checked on both occasions and the items checked on one of the two successive assessments and not checked on the other. Hence, items that were not checked at all, are not taken into account. The Jaccard's Similarity Ratio was first proposed by Jaccard (1908) in order to prevent two units from being considered similar because neither contains many of the attributes in a checklist. In the present study this ratio was used because it was expected that considering the generic character of the SIP68, a relatively large number of items would not be checked at all. Moreover, as the SIP68 assesses the impact that health status has on daily functioning, it is more interesting to know about the reliability of functional impacts that are present, than to know about impacts that are not present. Kappa, [15] a more widely known coefficient of agreement at item level, was not used as this coefficient considers items not checked on either occasion as an agreement. In a situation with a relatively large proportion of not applicable items, this would automatically result in a relatively high assessment of the level of agreement. Finally, the internal consistency of both lists was calculated and compared by means of Cronbach's α [16].

4.4. Results

The scoring procedure of the SIP136 calculates the score by attaching nominal weights to every item. In previous research we found that weighted and added scores (=score is the sum of the items checked) give approximately identical results [9,17]. Therefore, we used the more efficient procedure of added scores in scoring the short list. In table 4.3, apart from the mean scores on the short list and its separate categories at the four administrations, the total scores on the SIP136 at t3 and t4 are presented.
The total score on the sip136 was approximately 15. Using the SIP68, the highest scores are obtained on 'Mobility Control' and 'Social Behavior'. 'Emotional Stability' and 'Somatic Autonomy' appear to be slightly less affected than 'Psychic Autonomy & Communication' and 'Mobility Range'.

Table 4.3. Mean scores on the SIP68 and the SIP136

(n=51)	maximum possible score	t1 (st.dev.)	t2 (st.dev.)	t3 (st.dev.)	t4 (st.dev.)
total SIP68	68	**10.5** (8.0)	**10.6** (8.6)	**10.0** (8.0)	**9.5** (7.7)
somatic autonomy	16	**0.6** (1.5)	**0.7** (1.4)	**0.7** (1.4)	**0.6** (1.3)
mobility control	12	**4.0** (3.1)	**4.1** (3.5)	**4.1** (3.4)	**4.0** (3.1)
psych. autonomy & communication	11	**0.9** (1.5)	**0.8** (1.3)	**0.9** (1.5)	**0.9** (1.6)
social behavior	12	**3.6** (2.7)	**3.5** (2.6)	**3.1** (2.4)	**2.9** (2.5)
emotional stability	6	**0.5** (0.9)	**0.6** (1.0)	**0.3** (0.7)	**0.3** (0.7)
mobility range	10	**0.9** (1.8)	**1.0** (1.7)	**0.9** (2.2)	**0.7** (1.5)
SIP136	100			**15.5** (11.5)	**14.7** (10.7)

Table 4.4. Test-retest reliability of the SIP68 and its categories, expressed in Intraclass Correlation Coefficient

(n=51, p< .001)

	icc*
somatic autonomy	0.97
mobility control	0.95
psychic autonomy & communication	0.95
social behavior	0.94
emotional behavior	0.94
mobility range	0.90
sip68 totalscore	**0.97**

* icc = intraclass correlation coefficient

To estimate the test-retest reliability of the SIP68 (both extracted from the SIP136 and administered as a separate instrument), the ICC was calculated for the total instrument and for every category, using scores on the four administrations. Results are presented in table 4.4.

The ICC possible ranges from 0 to 1. Table 4.4. shows that the SIP68 and its categories has high to very high Intraclass Correlation Coefficients. These findings suggest good to very good test-retest reliability for the total list as well as for the separate categories. Data on the test-retest reliability at item level are presented in table 4.5.

Table 4.5. Test-retest reliability of the SIP68 at item level, expressed in Jaccard's Similarity Ratio

(n=51)

	jaccard's similarity ratio	
	mean SIP68	categories[1]
t1 - t2	0.60	0.53 - 0.67
t3 - t4	0.60	0.47 - 0.64
t1 - t3	0.42	0.35 - 0.57
t2 - t4	0.50	0.37 - 0.62

[1] minimum and maximum finding when the mean JSRs for every category is computed seperately

Table 4.6. Internal consistency of the SIP68

(n=51)

	cronbach's α	
	total	categories*
t1	0.90	0.54 - 0.83
t2	0.92	0.53 - 0.87
t3	0.90	0.49 - 0.85
t4	0.90	0.55 - 0.82

* lowest and highest α found for the separate categories.

The JSRs found suggest considerable agreement between both pairs of lists (t1-t2, t3-t4) at item level. Very little difference in the level of reliability is found between the two different administration types (t1-t2: SIP68 as a separate instrument, versus t3-t4: SIP68 extracted from the SIP136). Slightly less high JSRs are found when the agreement across the two administration types is calculated (t1-t3, t2-t4). All JSRs found, however, suggest considerable agreement in terms of items checked. Apart from test-retest reliability, the internal consistency of the instrument in both situations was calculated using Cronbach's α, which is the most common assessment of internal consistency. Findings are presented in table 4.6.

This table shows once more that there was no difference in reliability between the two administration types. Moreover, the internal consistency of the SIP68 and its categories appeared to be approximately of the same level as that of the original instrument [6].

4.5. Discussion and conclusion

The research subjects were 51 people with different rheumatic diagnoses. At t1 subjects were a sample from the population of a rheumatic outpatient clinic. A number of different diagnoses were represented. As rheumatic diseases are known to occur more often in women than in men, and more often in old age than in youth, the relative over representation of women and the mean age of 57 in our group is not surprising or disturbing. By excluding all respondents who obtained a score of zero on the SIP68, all respondents that judged themselves to be in (very) good health, and who were judged to have (very) good functional status by the rheumatologist, a more or less homogenous group of subjects was obtained with a functional status that was actually affected by a rheumatic disease. Although a relatively large percentage of the respondents at t1 did not complete the study (40%), no indications were found for a selective non-response that might do considerable harm to the generalizability of the findings in this study to the population of rheumatic patients

in general. External validity is supported by the score on the SIP136 found in our group. This score was approximately the same as found by Deyo et. al. in 1982 in a populations of rheumatoid arthritis patients: around 15%. In this study the reliability of a generic short version of the SIP was studied in two administration forms. The instrument was studied as an instrument on its own (SIP68), and as it was extracted from the administration of the original SIP (SIP136).
To establish the test-retest reliability the Intraclass Correlation Coefficients (ICC) and the Jaccard's Similarity Ratio (JSR) were computed for both administration types. Test-retest reliability coefficients for the total instrument and for the separate categories (ICC) ranged from 0.90 for category 'mobility range', to 0.97 for the total instrument. These high to very high coefficients suggest good to very good test-retest reliability of the SIP68. In the review on the SIP136 we performed some time ago [6], we did not find a report on ICC data of the SIP136. A comparison of the original SIP136 with the relatively new SIP68 on this characteristic, is therefore not possible. The test-retest data we did find were Pearson's correlation coefficients between two successive administrations of the SIP136. These ranged from 0.75 to 0.92 [6]. Deyo states that a test-retest correlation coefficient of 0.90 is 'desirable for instruments intended for use with individual patients or small groups' [19]. Hence, the short instrument in both administration types, judged by Intraclass Correlation Coefficients, is very reliable for individual as well as group use, as far as test-retest reliability is concerned.
The ICC, however, calculates reliability (or reproducibility) of scores on the total SIP68 or its categories. It might well be that in two successive administrations identical scores are obtained by checking the same number of items, but that the actual items checked differ greatly. In this respect the Jaccard Similarity Ratio (JSR) is a more accurate measure of reproducibility as it is a measure based on the actual congruence in individual items checked on successive occasions. It was already stated that the JSR is a very conservative measure of agreement as only items that are checked on at least one administration are taken into account. Items that are not checked on any administration, although consistently not checked, are not counted as agreements. In generic instruments like the SIP, where most patients only check a relatively small number of items, this is a relevant consideration as many 'double zero' items might (artificially) enhance the reliability assessment. The JSRs found within both administration types are equally high, and in agreement with agreement measures as reported in studies on the SIP136 [2,6,11]. The slightly less high JSRs which resulted when agreement was calculated not within but across administration forms, is probably due to the fact that at t3 and t4 the SIP136 was administered. This list contains twice the number of items of the SIP136. Apparently, respondents are consistent in the items they check if their choice has to be made from the same items (high JSR t1-t2 and t3-t4). However, if they are offered a broader choice at one time and less broad at the other (t1 versus t3, and t2 versus t4), they vary slightly more. However, the level of JSR found between t1 and t3 on the one hand and t2 and t4 on the other, still suggests considerable agreement between these administrations. From this it can be concluded that the scores on the SIP68 are stable and reliable, and that the descriptive profile provided by the SIP68 at item level, is also reliable and is not influenced by the way the instrument is administered.

A possible source of bias in test-retest reliability assessment is the fact that respondents remember the answers they gave the first time, when they fill out the list the second time. This might lead to an artificially enhanced assessment of the reliability. To prevent this form of bias, the respondents were asked to wait 48 hours between t1 and t2, and between t3 and t4. In the design of this study there is no possibility to determine the exact time the list was filled out. Supposing that respondents will have posted the return envelopes on the same day they filled out the list, the date of the postmarks on the return envelopes and the day of arrival of these envelopes provide an indication for the moment at which the lists were filled out. As no envelope arrived within the expected time and all postmarks were dated at least two days after t1 or t3 respectively, it is not to be expected that respondents did not wait 48 hours before filling out the list the second and fourth time. From this it can be concluded that it is not probable that a memory bias occurred.
Apart from reproducibility (test-retest), the internal consistency was examined. The internal consistency of the SIP68 appeared to be good. For both administration types, the Cronbach's α's are sufficient to high for the categories, and high for the total list. Apparently the internal consistency of the short instrument is not influenced by the difference in administration. Moreover the alphas found using the SIP68 appeared to be just as high as the alphas found in the literature on the reliability of the SIP136 [6].
The findings reported above are very encouraging: the SIP68 appears to be a reliable instrument. Moreover, its reliability is not substantially influenced by the way it is administered. However, as the SIP136 is a generic instrument, further study is needed in which the SIP68 is tested for its validity and reliability in different diagnostic groups. Nevertheless, the conclusion seems justified that the SIP68 can be regarded as a very promising candidate to become a more efficient alternative for the SIP136.

Acknowledgment

We are grateful for the kind and useful cooperation of dr. B.A. Masek, rheumatologist in the 'st. Maartens Gasthuis' in Venlo, mrs. Dominique Tilli who assisted us in the data collection, and the respondents who made it possible to complete this study.

Appendix

SIP68

Somatic autonomy

1. I get around in a wheelchair.
2. I get dressed only with someone's help.
3. I do not move into or out of bed by myself, but am moved by a person or mechanical aid.
4. I stand up only with someone's help.
5. I do not fasten my clothing, for example require assistance with buttons, zippers, shoelaces.
6. I do not walk at all.
7. I do not use stairs at all.
8. I make difficult moves with help, for example, getting into or out of cars, bathtubs.
9. I do not bathe myself completely, for example, require assistance with bathing.
10. I do not bathe myself at all, but am bathed by someone else.
11. I do not have control of my bladder.
12. I am very clumsy in body movements.
13. I do not have control of my bowels.
14. I feed myself with help from someone else.
15. I do not maintain balance.
16. I use bedpan with assistance.
17. I am in a restricted position all the time.

Mobility control

1. I go up and down stairs more slowly, for example, one step at a time, stop often.
2. I walk shorter distances or stop to rest often.
3. I walk more slowly.
4. I use stairs only with mechanical support, for example , handrail, cane crutches.
5. I walk by myself but with some difficulty, for example, limp, wobble, stumble, have stiff leg.
6. I kneel, stoop or bend down only by holding on to something.
7. I do not walk up or down hills.
8. I get in and out of bed or chairs by grasping something for support, or using a cane or walker.
9. I stand only for short periods of time.
10. I dress myself, but do so very slowly.
11. I have difficulty doing handwork, for example turning faucets, using kitchen gadgets, sewing, carpentry.
12. I move my hands or fingers with some limitation or difficulty.

Psychic autonomy and communication

1. I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things.
2. I have difficulty doing activities involving concentration and thinking.
3. I react slowly to things that are said or done.
4. I make more mistakes than usual.
5. I do not keep my attention on any activity for long.
6. I forget a lot, for example, things that happened recently, where I put things, appointments.
7. I am confused and start several actions at a time.
8. I do not speak clearly when I am under stress.
9. I have difficulty speaking, for example, get stuck, stutter, stammer, slur my words.
10. I do not finish things I start.
11. I am having trouble writing or typing.

Social behavior

1. My sexual activity is decreased.
2. I am cutting down the length of visits with friends.
3. I am drinking less fluids.
4. I am doing fewer community activities.
5. I am doing fewer social activities with groups of people.
6. I am going out for entertainment less often.
7. I stay away from home only for brief periods of time.
8. I am eating much less than usual.
9. I am not doing heavy work around the house.
10. I do my hobbies and recreation for shorter periods of time.
11. I am doing less of the regular daily work around the house than I would usually do.
12. I am cutting down on some of my usual inactive recreation and pastime, for example, watching TV, playing cards, reading.

Emotional stability

1. I often act irritable toward those around me, for example, snap at people, give sharp answers, criticize easily.
2. I act disagreeable to family members, for example, I act spiteful, I am stubborn.
3. I have frequent outbursts of anger at family members, for example, strike at them, scream, throw things at them.
4. I act irritable and impatient with myself, for example, talk badly abut myself, swear at myself, blame myself for things that happen.
5. I am not joking with family members as I usually do.
6. I talk less with those around me.

Mobility range

1. I am not doing any of the shopping that I would usually do.
2. I am not going into town.
3. I am not doing any of the house cleaning that I would usually do.
4. I am not doing any of the regular work around the house that I would usually do.
5. I stay home most of the time.
6. I am not doing any of the clothes washing that I would usually do.
7. I am not going out to visit people at all.
8. I am getting around only within one building.
9. I have given up taking care of personal or household business affairs, for example paying bills, banking working on budget.
10. I do not get around in the dark or in unlit places without someone's help.

Literature

1. Bergner M, RA Bobbit, WB Carter, BS Gilson. The sickness impact profile: development and final revision of a health status measure. Med. Care 1981, 19: 787-805.
2. Pollard WE, RA Bobbit, M Bergner, DP Martin, BS Gilson. The Sickness Impact Profile: reliability of a health status measure. Med. Care 1976, 14: 146-155.
3. Bergner M, RA Bobbit, WE Pollard, DP Martin, BS Gilson. The Sickness Impact Profile, validation of a health status measure. Med. Care 1976,14: 57-67.
4. Deyo RA, TS Iniu, J Leininger, S Overman. Physical and psychosocial function in rheumatoid arthritis. Clinical use of a self-administered health status instrument. Arch. Intern. Med. 1992, 142: 879-882.
5. de Witte LP, H Jacobs, FG van der Horst, A Luttik, J Joosten, H Philipsen. De waarde van de Sickness Impact Profile als maat voor het functioneren van patinten. Gezondheidszorg en samenleving 1987, 9: 120-127.
6. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks. Sickness Impact Profile: the state of the art of a generic functional status measure. Soc. Sci. Med. 1992, 8: 1003-1014.
7. McDowell I, N Newell. Measuring Health: A Guide to Rating Scales and Questionnaires. Oxford University Press, New York, 1987.
8. Bowling A. Measuring health: a review of quality of life measurement scales. Open University Press, Buckingham, 1991.
9. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen. The Development of a short generic version of the sickness impact profile, J.Clin.Epid. 1994, 47, no.4: 407-418.
10. Kerlinger FN. Foundations of behavioral research. Holt, Richart and Winston, 1973.
11. Pollard WE, RA Bobbit, M Bergner. Examination of variable errors of measurement in a survey-based social indicator. Social Indicators Research 1987, 5: 279-301.
12. Bravo G, L Potvin. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. J.Clin.Epid. 1991, vol. 44: 381-390.
13. Jaccard P. Nouvelles recherches sur la distribution florale. Bull.Soc.Vaud.Sc. Nat. XLIV. 1908, 163: 223-270.
14. Dormaar JMM. Consensus in psychotherapy. Thesis University of Limburg, Maastricht, 1990.
15. Cohen J. A coefficient of agreement for nominal scales. Educational and Psycological measurement, 20, 37-46, 1960.
16. Cronbach LJ. Coefficient alpha and the internal structure of a test. Psychometrika 1951, 16: 297-334.
17. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks. Sickness Impact Profile: can a famous functional status measure be improved? 3rd. European Health Services Research Meeting. 1991 13-14 dec. London.
18. Deyo RA, TS Inui, J Leininger, S Overman. Physical and psychosocial function in rheumatoid arthritis: clinical use of a self-administered health status instrument. Arch.intern.Med. 1982, 142: 879-882.
19. Deyo R. Measuring Functional Outcomes in Therapeutic Trials for Chronic Disease. Controlled Clinical Trials 1984, 5: 223-240.

Chapter 5

The SIP68: A measure of health-related functional status in rehabilitation medicine

M.W.M. Post[1], A.F. de Bruin[2], L.P. de Witte[2,3], G. Schrijvers[1].

[1] University of Utrecht, Department of Medicine, Utrecht, The Netherlands
[2] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands
[3] IRV, Hoensbroek, The Netherlands

This paper was excepted for publication in The Archives of Physical Medicine and Rehabilitation

Abstract

Objective: To demonstrate the usefulness of the SIP68, a recently developed short version of the SIP, for measuring health related functional status in rehabilitation medicine. Design: Survey, oral interviews. Setting: Patient's homes. Patients: 315 persons (out of 423 that could be reached) with a spinal cord injury whose mean average age was 39.4 years and who were living in the community at the time of the interview.
Main outcome measures: Internal consistency is tested by computing Cronbach's α. Construct validity is tested by principal components analysis and computing Cattell's similarity index. Criterion validity is tested by comparing SIP68 results with the level of the spinal cord lesion and with those of specific measures of disability (Barthel Index), and life satisfaction (Life Satisfaction Scale), and with vocational status.
Results: SIP68 scores and subscale scores indicate that our spinal cord injured group falls well within the scope of this instrument. Internal consistency figures are good and the proposed six-dimensional structure is confirmed. Criterion validity figures are also satisfactory. Barthel Index scores show high agreement with the scores of the subscale Somatic Autonomy, moderate agreement with the other physically and with the socially related subscales, and low agreement with the mentally related subscales of the SIP68. LSS scores show low agreement with the physically related subscales and moderate agreement with the mentally and socially related subscales. Figures of vocational status show strongest agreement with the socially oriented subscales. Conclusion: The SIP68 is recommended as a useful generic outcome measure for research in rehabilitation medicine.

5.1. Introduction

The Sickness Impact Profile (SIP) was first published in 1976. Since its revision in 1981, it has become one of the best known generic health status measures [1,2]. The SIP is a measure of 'health-related changes in behaviour associated with the carrying out of one's daily activities'. The SIP consists of 136 items. Every item describes a possible sickness impact. The instrument has been highly recommended [3-6] and was even advised as a criterion standard [7]. A recent review stated that the broad coverage of the SIP, its extensive development and its emphasis on current performance makes it a candidate for selected applications in rehabilitation [8]. In 1985, a Dutch version of the SIP was published and proved to have good validity and reliability [9,10].
Compared to related instruments, the SIP is very long. A recent Dutch study with patients with stroke [11] found that it was time consuming and tiring for patients in early recovery. Deyo found that respondents experienced items not applying to their situation as redundant [12]. Several authors state its length is an obstacle to routine use [5,8]. Therefore, a short generic version of the SIP was recently developed: the SIP68 [13]. Evaluation of this short instrument during its development revealed good reliability and validity in a multi-diagnostic population and in a population of patients with rheumatiod arthritis. The SIP68 is meant to be a generic measure, hence its psychometric properties have to be good in populations from any diagnostic group. In this paper, therefore, the SIP68 will be validated as a measure of health-related functional status in a group of patients with a spinal cord injury.
In this paper, the SIP68 will first be briefly introduced. To clarify the relevance of the instrument for rehabilitation research, the position of the SIP68 in relation to the International Classification of Impairments, Disabilities and Handicaps (ICIDH) is discussed. After this, the study population and methods will be presented. In the 'Results' paragraph, findings concerning reliability and validity will be discussed.

5.2. The SIP68

The SIP68 contains 68 items, which are all statements regarding behaviour (Appendix). Respondents are asked to check those items that both apply to their situation on the day they fill out the list and that are related to their health status. The items are divided over 6 subscales. In every subscale a higher score indicates more health-related behavioural problems. The first subscale, 'somatic autonomy' (SA), deals with basic somatic functions, like getting dressed, standing, walking, eating, and the fact that help is needed with these functions. Items in the second subscale, 'mobility control' (MC), describe behaviour that is related to the level of control which an individual has over his or her body. Items in this subscale are related to walking, or have to do with hand and arm control. 'Psychic autonomy and communication' (PAC), the third subscale, describes behaviour that is associated with the level at which an individual is able to operate without help in areas of mental functioning, including the (verbal) ability to communicate. The possible

consequences of a health deviation in a person's functioning in relation to other people is described in the subscale 'social behaviour' (SB). Sexual activity, visiting friends, and activities in groups of people are included here. The fifth subscale is 'emotional stability' (ES). It assesses the effect health status has on the emotional status of a respondent. The 6 items are concerned with irritability, and/or acting disagreeably with oneself or others. The final subscale, 'mobility range' (MR), is concerned with the range of actions to wich a person has (limited) disposition, given his or her health status, like shopping, housecleaning and taking care of personal business affairs.
The selection of items of the SIP68 was made on the basis of a multidiagnostic database containing over 830 completed SIP's [13]. First, the subscale 'work' and items that were hardly ever checked were removed from the list. Further selection was based on findings from principal components analyses of the remaining items. This resulted in a 68-item version of the SIP on six subscales. The six subscales together were able to predict total SIP136-scores almost perfectly ($R^2 = 0.96$) [13]. Further evaluation of the SIP68 was done by administrating the SIP68 four times to 51 persons with rheumatic health problems. In this population, the internal consistency of the SIP68 proved to be good, with Chronbach's α for the total scale on seperate administrations between 0.90 and 0.92, and for individual subscales between 0.49 and 0.87. Test-retest reliability also proved to be good, the Intra Class Correlation Coefficient was 0.97 for the total score and between 0.90 and 0.97 for individual subscales [14].

In this study, the standard format and scoring procedures of the SIP68 are replicated, with one important exception. On SIP136 and SIP68 items only two answers are possible (applies or does not apply to my situation). This may cause deceptive figures in a population that includes many wheelchair dependent people, because people who cannot walk at all are forced to score negatively on all items relating to difficulties with walking (for example 'I walk shorter distances or stop to rest often') and hence obtain a low score, wrongly indicating good health related status. For the SIP136, item weights are said to be the solution to this problem. This is not likely to be sufficient, because of the doubtful effects of weighting in general as described by Jenkinson [15]. Moreover, in the case of the SIP136, weighted scores and 0/1 scores almost perfectly agreed (correlations between 0.94 and 0.97 for subscale and total scores) [13]. We solved this problem by scoring positively all items relating to difficulties with walking, when the item 'I do not walk at all' was scored positively. This procedure affects one item in the subscale SA and seven items in the subscale MC (given an asterix in appendix A). This method is now entered in the standard SIP68 administration guidelines [16].

5.2.1. SIP68 and the ICIDH

Within the framework of the International Classification of Impairments Disabilities and Handicaps (ICIDH) [17], the concept that the SIP68 aims at is closely related both to disability and to handicap. Some subscales in the SIP68 (eg. the subscales

SA and MC) can be considered as a description of the level of disability. Other subscales describe role performance (eg. the subscales MR and SB), thus being a part of the description of handicap. The WHO-definitions of disability and of handicap, however, refer to a general level of ability considered normal for an individual. The SIP68 does measure restrictions in behaviour, but the respondent is asked to refer to his own normal level and pattern of behaviour. Hence, not all restrictions on the level of disability or handicap that might follow from a certain state of health (or impairment) are represented in a SIP68-score; only those restrictions that are relevant in this respondent's situation, given his/her premorbid behaviour are represented. From this the SIP68 can be said to measure individualized levels of disability and handicap. This is similar to the way in which these concepts are often used in formulating the goals of rehabilitation medicine [8,18]. Hence, it may be concluded that the SIP68 is a very relevant instrument to be used in rehabilitation medicine outcome research.

5.3. Population

For this study individuals with a Spinal Cord Injury (SCI) who were between 18 and 65 years of age, and who were rehabilitated after injury in a specialized rehabilitation centre between 1986 and 1992, were asked to participate. 525 persons were available. From this group 423 persons could be reached, and 315 persons participated in the study (response rate 60%). There were no statistically significant differences at an alpha level of 1% between the response group and the non-response group with regard to type of SCI, cause of SCI, age, sex and time after discharge from the rehabilitation centre. All respondents were interviewed at home. In the response group, the mean age is 39.4 years (sd 12.5). The majority are male (75.4%) and 57.1% are married or living together. The types of SCI include 21.7% complete quadriplegia, 20.4% incomplete quadriplegia, 29.2% complete paraplegia and 28.6% incomplete paraplegia. The main causes of SCI in the response group are traffic accidents (34.9%), followed by 15.1% sports accidents, 12.9% occupational accidents, 10.4% resulting of bodily processes, 9.1% falls, and 7.5% resulting from medical treatment. Time after injury is between 1 and 7 years (mean 3.6 years, sd 1.9). Most respondents (60.0%) are wheelchair dependent.

5.4. Instruments

The level of injury is scored by asking respondents and checking the answer in available medical files. Analyses with this variable are restricted to individuals with a complete SCI, because in the case of incomplete lesions the amount of damage is more important in determining disability than the level of the injury. For measuring physical disability the Barthel Index (BI) is used as criterion standard. Within its scope, the Barthel Index is one of the best instruments available for use in a group of physically disabled [5,7]. We use the 10-item version described by Wade and Collin

[19-21] that has already been used in many studies and has proven to have good reliability and validity [5]. A Dutch translation of this BI version has a Cronbach's α of 0.96 and an inter-rater reliability (Kappa) of 0.88 [22]. We adapted this BI-version for use in an interview situation, wich may be considered to lead to comparable results [23]. A more comprehensive analysis of the clinimetric properties of our BI version will be reported elsewhere [24]. BI scores are between zero and 20, reflecting complete dependence to complete independence in self-care and mobility (mean score in this study 13.6, sd 5.4; Cronbach's α 0.87). Social functioning is measured here as being vocationally active (having a paid job, studying or housekeeping). According to this criterion 150 persons (47.6%) are vocationally active and 165 (52.4%) are not.

Life satisfaction is measured by the Life Satisfaction Scale [25]. The LSS measures satisfaction with life in general and with eight specific domains (for example self-care and family relations). LSS-item scores range from 1 (very dissatisfying) to 6 (very satisfying). A total LSS-score is computed by adding all nine item scores and dividing them by nine (range in this study 1.9-6, mean 4.3, sd 0.8; Cronbach's α 0.76).

5.5. Statistical procedures

The reliability of the SIP68 is tested by computing Chronbach's α. This is a measure for internal consistency, or the degree in wich items in a scale are related to each other. Chronbach's α has a possible range from 0 to 1. Scores above 0.53 are considered to be satisfactory, and scores above 0.80 are considered to indicate good internal consistency [26]. Construct validity of an instrument is concerned with the level to which that instrument really measures the construct under study. It can be assessed by studying the extent to which a measure relates to other measures, in accordance with theoretically derived hypotheses about concepts that are being measured [27]. This is the external view on construct validity, and is here referred to as criterion validity. The internal view on construct validity is to investigate whether the theoretically expected dimensions of the concept (and their inter-relations) can be demonstrated in empirical data. When this structure is found, the construct validity of the instrument is supported. The SIP68 consists of six subscales that represent different elements of health-related functional status and therefore are supposed to have low correlations with each other. Subscales representing related concepts (SA and MC, MC and SB, ES and PAC) are supposed to have higher correlations with each other than subscales measuring unrelated concepts (eg SA and PAC). A more formal way to validate the internal structure of the concept measured, is by confirming the proposed structure in the data collected by the instrument under study. Cattell's similarity index s is a measure of agreement in the pattern of factor loadings between a proposed structure and the structure found in empirical data. It is calculated by comparing the obtained factor loadings of the instrument's items with the proposed factor structure. Based on the level of similarity found, a p-value is obtained for the possibility of finding this level of similarity by chance [28].

Criterion validity tests are performed by relating SIP68 scores to BI scores, to LSS scores and to social functioning. To pass this criterion validity test, subscales that measure related concepts to these criterion standards, for example BI and SA, LSS and ES, social functioning and SB, should show correlations that are (a) statistically significant (b) substantial (> 0.30), and (c) are higher than the correlations of subscales that measure concepts not related with the critera. All analyses are performed with SPSS/PC+ and as far as possible non-parametric tests are used: Spearman-Brown correlations and Mann-Whitney order of ranks tests for differences between groups. A p-value lower than 0.01 is considered to be statistically significant.

5.6. Results

5.6.1. Frequencies

Table 5.1 presents mean scores on the SIP68 and each individual subscale. In this diagnostic group, the mean SIP68 score is 22.8, or about 30% of the theoretical maximum score. Only very few respondents reach a maximum score on one or more subscales. Scores on the subscales ES and PAC are strongly negatively skewed, with median scores of 0.

5.6.2. Internal consistency

Cronbach's α of the SIP68 is good. Five out of six subscales also have good reliability (0.72 to 0.80), only one subscale, ES, lags behind but its reliability is still acceptable (0.68).

Table 5.1. Frequency of scores in the SIP68 and individual subscales

	number of items	mean score	standard deviation	% zero scores	% maximum scores
somatic autonomy	17	5.8	4.8	18.0	0.6
mobility control	12	7.4	2.4	3.2	1.6
mobility range	10	2.4	2.4	31.1	1.6
social behaviour	12	5.1	3.1	6.0	1.6
emotional stability	6	1.0	1.4	54.0	2.5
psych. aut. & comm	11	1.1	1.9	53.8	0.3
total SIP68	68	22.8	11.1	0.3	0.0

Table 5.2. Correlations between subscales and with SIP68

	S.A.	*M.C.*	*M.R.*	*S.B.*	*E.S.*	*P.A.C.*
somatic autonomy	xxx					
mobility control	.54	xxx				
mobility range	.54	.45	xxx			
social behaviour	.41	.42	.67	xxx		
emotional stability	.12	.08	.27	.48	xxx	
psych. aut. & comm.	.21	.26	.35	.47	.43	xxx
SIP68	.80	.67	.79	.80	.44	.53

Note: all correlations > .20 are statistically significant ($p < 0.001$).

5.6.3. Construct validity

In table 5.2 correlations between the subscales are presented. The intercorrelations range from 0.08 up to 0.67. With one exeption (MR with SB) these correlations are not very high, so little redundancy seems to be present. Correlations between conceptually related subscales (SA with and MC, MC with SB, ES with PAC) are higher than the correlations between conceptually related subscales. The two socially related subscales, MR and SB, have substantial correlations with all other subscales, perhaps indicating that a low score on these subscales presupposes good physical as wel as good mental functioning. To investigate the impact of our adaptation of the scoring procedure, we also computed correlations between subscales when following the original applying/not applying method. Our scoring procedure substantially improves construct validity. For example, the correlation of MC with SA, two subscales supposed to be positively related to each other, changes as a result of the scoring procedure from -0.42 into 0.54.
At item level, principal components analysis also confirms the proposed structure of the SIP68. Cattell's salient similarity index s shows good similarity figures between the proposed factor structure and the obtained factor structure for all subscales.

Table 5.3. Similarity of obtained factor solution with proposed factor solution

	SA	MC	PAC	SB	ES	MR
s	.63	.57	.95	.37	.72	.55
vS	.42	.39	.42	.34	.51	.42
p	.000	.000	.000	.002	.000	.000

s: Cattell's salient similarity index s; vS: criterion for similarity, if s exeeds vS than the factors are reliably similar; p: estimate of probability values for s.

Table 5.4. Correlations of SIP68 and individual subscales with level of lesion BI and LSS

	Level of lesion (N=158)	Barthel Index (N=315)	Life Satisfaction (N=315)
somatic autonomy	-0.72*	- 0.91*	- 0.32*
mobility control	-0.47*	- 0.47*	- 0.22*
mobility range	-0.39*	- 0.54*	- 0.42*
social behaviour	-0.29*	- 0.41*	- 0.53*
emotional stability	-0.07	- 0.11	- 0.41*
psych. aut. & comm.	-0.31*	- 0.21*	- 0.34*
SIP68	-0.59*	- 0.74*	- 0.52*

* = $p < 0.001$

5.6.4. Criterion validity

A test on criterion validity is made by relating SIP68 scores to the level of the lesion, and to the BI, the LSS and to being vocationally active. The SIP68 as well as 5 out of 6 subscales show a significant relation with the location of lesion. ES does not seem to be related to the severity of the injury. With regard to the subscale PAC, inspection of item frequencies reveals that a slightly higher score of cervical injured on this subscale is mainly caused by positive scores on the item 'I am having trouble writing or typing', which in this diagnostic group is probably the result of poor hand muscle control instead of mental problems.
Table 5.4 also shows a high correlation between the BI and the SIP68 score. This correlation is negative, because a high BI-score indicates physical independence, wereas a high SIP68 score indicates a poor health-related functional status. The subscale SA alone almost exactly predicts BI-scores, explaining 82.8% of all BI-variance. As expected, correlations of BI with ES and PAC are considerably lower, although they are still significant at the 1% level.
The LSS-score is related most strongly to the SIP68 total score and to the subscales related to social and to mental functioning. Surprisingly, the subscale showing the highest correlation with the LSS is SB and the PAC subscale is only moderately related to the LSS score. The 0.52 correlation of the total SIP68 score with the LSS scores indicates substantial agreement between both instruments. SIP68 scores also are related to being vocationally active: persons being vocationally active have statistically significant lower SIP68-scores, particularly for MR, SB, SA and for the total score.

Table 5.5. Relationship of SIP68 scores with vocational activity

	not vocationally active (N=165)	vocationally active (N=150)	value of z in Mann-Whitney test
somatic autonomy	7.04	4.53	- 4.63**
mobility control	7.86	6.85	- 3.89**
mobility range	3.06	1.57	- 5.36**
social behaviour	5.85	4.29	- 4.62**
emotional stability	1.08	0.89	- 1.47
psych. aut. & comm.	1.40	0.87	- 3.21*
SIP68	26.30	19.00	- 5.61**

* = $p < 0.01$; ** = $p < 0.001$

5.7. Discussion

In this study, the usefulness of the SIP68 for rehabilitation outcome research is investigated. On theoretical grounds it appears to be a relevant instrument. According to the raw scores obtained, our group of spinal cord injured falls well within the scope of the SIP68. However, persons who live in a rehabilitation centre or nursing home and SCI-persons above 65 years of age were excluded from the study. These groups might have obtained higher SIP68 scores, but mean scores on all subscales indicate that these groups would probably also have fallen within the scope of the instrument. Figures about the alpha reliability of the SIP68 and all subscales are satisfactory. Unfortunately it was not possible to obtain a test-retest reliability score, but the results of previous research on the SIP68 may justify some confidence on this point [14]. The proposed six-dimensonal structure of the SIP68 is confirmed by Cattel's similarity index. In addition evidence concerning the criterion validity of the SIP68 is presented. The expected pattern of relationships with the level of the lesion, and with measures of physical disability, vocational activity and mental functioning is reproduced to a satisfactory degree .

The very high correlation between SA and the BI indicates that in using the SIP68 there is no need for additional instruments regarding self care ability like the Barthel Index. The SIP68, however, also covers a broader concept. SIP68 scores are related to LSS scores to a considerable degree, but they do not cover the LSS completely. Perhaps this is due to conceptual differences between health related functional status as measured by the SIP68 [29], and life satisfaction.

Eight questions about difficulties with walking were recoded in order to avoid deceptive results. This substantially influenced scores on the subscale MC, and resulted in better reliability and validity results. This recoding is now added to the standard administration procedure for the SIP68.

It is not yet possible to compare our results to those of other research. Experience with the SIP136 in spinal cord injury populations is also rare [30-35]. McColl and Rosenthal used an adaptation of the psychosocial subscale of the SIP (16 items), and reported an α reliability of 0.77, a mean score of 2 and 30% zero scores [33]. These results are compararable to ours, if we treat ES and PAC together as one scale (17 items, mean score 2.1, Cronbach's α 0.80). Lundqvist at al report significant differences in overall SIP score and physical dimension score between tetraplegics and paraplegics, and a strong correlation between degrees of neurologic deficits and physical dimension scores. Scores reflecting psychosocial functions did not discriminate between neurologic subgroups or Frankel grades [31]. These results also are comparable with ours. McColl and Rosenthal reported a slightly higher correlation between psychosocial SIP and Life Satisfaction than we did (-0.51, against -0.41 for ES and -0.34 for PAC in this research), but used a different measure of Life Satisfaction, so this difference cannot lead to any conclusions. We conclude form the scarce evidence that the SIP68 in this population is comparable to with the much longer original SIP.

This paper illustrates the usefulness of the SIP68 as a generic measure for outcome research in rehabilitation medicine. The results about reliability and validity reported here add to those reported by De Bruin at all [13,14]. Further research should bring additional evidence and compare the usefulness of the SIP68 with that of other generic measures of functional health status.

Appendix

SIP68

Somatic autonomy

1. I get around in a wheelchair.
2. I get dressed only with someone's help.
3. I do not move into or out of bed by myself, but am moved by a person or mechanical aid.
4. I stand up only with someone's help.
5. I do not fasten my clothing, for example require assistance with buttons, zippers, shoelaces.
6. I do not walk at all.
7. I do not use stairs at all.
8. I make difficult moves with help, for example, getting into or out of cars, bathtubs.
9. I do not bathe myself completely, for example, require assistance with bathing.
10. I do not bathe myself at all, but am bathed by someone else.
11. I do not have control of my bladder.
12. I am very clumsy in body movements.
13. I do not have control of my bowels.
14. I feed myself with help from someone else.
15. I do not maintain balance.
16. I use bedpan with assistance.
17. I am in a restricted position all the time.

Mobility control

1. I go up and down stairs more slowly, for example, one step at a time, stop often.
2. I walk shorter distances or stop to rest often.
3. I walk more slowly.
4. I use stairs only with mechanical support, for example , handrail, cane crutches.
5. I walk by myself but with some difficulty, for example, limp, wobble, stumble, have stiff leg.
6. I kneel, stoop or bend down only by holding on to something.
7. I do not walk up or down hills.
8. I get in and out of bed or chairs by grasping something for support, or using a cane or walker.
9. I stand only for short periods of time.
10. I dress myself, but do so very slowly.
11. I have difficulty doing handwork, for example turning faucets, using kitchen gadgets, sewing, carpentry.
12. I move my hands or fingers with some limitation or difficulty.

Psychic autonomy and communication

1. I have difficulty reasoning and solving problems, for example, making plans, making decisions, learning new things.
2. I have difficulty doing activities involving concentration and thinking.
3. I react slowly to things that are said or done.
4. I make more mistakes than usual.
5. I do not keep my attention on any activity for long.
6. I forget a lot, for example, things that happened recently, where I put things, appointments.
7. I am confused and start several actions at a time.
8. I do not speak clearly when I am under stress.
9. I have difficulty speaking, for example, get stuck, stutter, stammer, slur my words.
10. I do not finish things I start.
11. I am having trouble writing or typing.

Social behavior

1. My sexual activity is decreased.
2. I am cutting down the length of visits with friends.
3. I am drinking less fluids.
4. I am doing fewer community activities.
5. I am doing fewer social activities with groups of people.
6. I am going out for entertainment less often.
7. I stay away from home only for brief periods of time.
8. I am eating much less than usual.
9. I am not doing heavy work around the house.
10. I do my hobbies and recreation for shorter periods of time.
11. I am doing less of the regular daily work around the house than I would usually do.
12. I am cutting down on some of my usual inactive recreation and pastime, for example, watching TV, playing cards, reading.

Emotional stability

1. I often act irritable toward those around me, for example, snap at people, give sharp answers, criticize easily.
2. I act disagreeable to family members, for example, I act spiteful, I am stubborn.
3. I have frequent outbursts of anger at family members, for example, strike at them, scream, throw things at them.
4. I act irritable and impatient with myself, for example, talk badly abut myself, swear at myself, blame myself for things that happen.
5. I am not joking with family members as I usually do.
6. I talk less with those around me.

Mobility range

1. I am not doing any of the shopping that I would usually do.
2. I am not going into town.
3. I am not doing any of the house cleaning that I would usually do.
4. I am not doing any of the regular work around the house that I would usually do.
5. I stay home most of the time.
6. I am not doing any of the clothes washing that I would usually do.
7. I am not going out to visit people at all.
8. I am getting around only within one building.
9. I have given up taking care of personal or household business affairs, for example paying bills, banking working on budget.
10. I do not get around in the dark or in unlit places without someone's help.

Literature

1. Bergner M, RA Bobbit, WB Carter, BS Gilson. The sickness impact profile: development and final revision of a health status measure. Med. Care 1981, 8:787-805.
2. Bergner M. Development, testing and use of the Sickness Impact Profile. In: Walker, S.R. and R.M. Rosser. (Ed) Quality of life Assessment. Dordrecht. Kluwer, 1993.
3. Spitzer WO. Keynote address: State of science 1986: quality of life and functional status as target variables for research. J. Chron. Dis. 1987, 40: 465-474.
4. Bowling A. Measuring health: a review of quality of life measurement scales. Philadelphia. Open University Press, 1991.
5. Wilkin D, L Hallam, MA Doggett. Measures of need and outcome for primary health care. Oxford. Oxford University Press 1992.
6. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks. Sickness Impact Profile: the state of the art of a generic functional status measure. Soc. Sci. Med. 1992, 35: 1003-14.
7. Mcdowell I, C Newell. Measuring Health: A guide to rating scales and questionnaires. New York: Oxford University Press, 1987.
8. Keith RA. Functional status and health status. Arch. Phys. Med. Rehabil. 1994, 75: 478-83.
9. Luttik A, HM Jacobs, LP de Witte. De Sickness Impact Profile (Dutch translation), 1987.
10. de Witte LP, H Jacobs, F van der Horst, A Luttik, J Joosten, H Philipsen. De waarde van de Sickness Impact Profile als maat voor het functioneren van patiënten. Gezondheid en Samenleving 1987, Jrg 8, 2: 120-127.
11. Schuling J, J Greidanus, B Meijboom-De Jong. Measuring functional status of stroke patients with the Sickness Impact Profile. Disabil. Rehab. 1993, 15: 19-23.
12. Deyo RA. Measuring functional outcomes in therapeutical trials for chronic disease. Contr. Clin. Trials 1984, 5: 223-240.
13. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen. The development of a short generic version of the Sickness Impact Profile. J. Clin. Epidem. 1994, 47: 407-418.
14. de Bruin AF, M Buys, LP de Witte, JPM Diederiks. The Sickness Impact Profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. J. Clin. Epid. 1994, vol. 47, no.8: 863-871.
15. Jenkinson C. Why are we weighting? A critical examination of the use of item weights in a health status measure. Soc. Sci. Med. 1991, 32: 1413-1416.
16. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen. SIP68, a shortened version of the Sickness Impact Profile. Internal publication, university of Limburg, Dept. Medical Sociology. Maastricht 1995.
17. World Health Organisation. International classification of impairments, disabilities and handicaps. Geneva 1980.
18. Dijk AJ van. Revalidatiegeneeskunde: theorie en praktijk. Med. Contact 1992, 47: 1421-1423.
19. Wade DT, C Collin. The Barthel ADL Index: a standard measure of physical disability? Int. Disab. Stud. 1988, 10: 64-67.
20. Collin C, DT Wade, S Davies, V Horne. The Barthel ADL Index: a reliability study. Int. Disabil. Stud. 1988, 10: 61-3.
21. Wade DT, RL Hewer. Functional abilities after stroke: measurement, natural history and prognosis. J. Neurol. Neurosurg. Psychiatr. 1987, 50: 177-82.
22. Haan R de, M Limburg, J Schuling, J Broeshart, L Jonkers, P van Zuylen. Klinimetrische evaluatie van de Barthel Index, een maat voor beperkingen in het dagelijks functioneren. Ned. Tijdschr. Geneeskd. 1993, 137: 917-21.
23. Shinar D, CR Gross, KS Bronstein, et al. Rehability of the activities of daily living scale and its use in telephone interview. Arch. Phys. Med. Rehabil. 1987, 68: 723-728.
24. Post MWM, FWA van Asbeck, AJ Dijk, AJP Schrijvers. Een nederlandse interview-versie van de Barthel Index (a Dutch version of the..). Ned. Tijdschr. Geneesk. 1995, in press.
25. Bränholm IB, M Ecklund, KS Furgl-Meyer, AR Furgl-Meyer. On work and Life satisfaction. J. Rehab. Sc. 1991, 4 (2): 29-34.
26. Nunnally JC. Psychometric theory. New york, McGraw, 1967.

27. Carmines EG. RA Zeller. Reliability and validity assessment. Beverly Hills/London Sage publ. 1979.
28. Tabachnick BG, LS Fidell. Using multivariate statistics, 2nd edn. New York: Harper & Row 1989;642-644.
29. Nydevik I, K Hulter-Asberg. Sickness Impact after Stroke. A 3-year follow-up. Scand.J.Prim. Health Care 1992: 10: 284-9
30. Siösteen A, C Lundqvist, C Blomstrand, L Sullivan, M Sullivan. The quality of life of three functional spinal cord injury subgroups in a swedish community. Paraplegia 1990, 28: 467-488.
31. Lundqvist C, A Siösteen, C Blomstrand, B Lind, M Sullivan. Spinal Cord Injuries. Clinical, functional and emotional status. Spine 1991, 16: 78-83.
32. de Witte LP. After the rehabilitation centre a study into the course of functioning after discharge from rehabilitation (dissertation) Amsterdam/Lisse. Zwets en Zeitlinger. 1992.
33. McColl MA, C Rosenthal. A model of resource needs of aging spinal cord injured men. Paraplegia 1994, 32: 261-270.
34. Elliott TR, SM Herrick, TE Witty, F Godshall, M Spruell. Social relationships and psychosocial impairment of persons with spinal cord injury. Psychology and Health 1992, 7: 55-67.
35. Richards JS, FJ Osuna, TM Jaworski, TA Novack, DA Leli, TJ Boll. The Effectiveness of Different Methods of Defining Traumatic Brain Injury in Predicting Postdischarge Adjustment in a Spinal Cord Injury Population. Arch. Phys. Med. Rehabil. 1991, 72: 275-279.

Chapter 6
Assessing the responsiveness of a functional status measure: the sickness impact profile versus the SIP68

A.F. de Bruin[1], J.P.M. Diederiks[1,2], L.P. de Witte[1,2], F.C.J. Stevens[1], H. Philipsen[1].

[1] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands
[2] IRV, Hoensbroek, The Netherlands

This paper has been submitted for publication

Abstract

In this study the Sickness Impact Profile and the SIP68 are studied for their ability to detect changes in health related behavioral status. Methodological approaches towards responsiveness are inventarized and discussed. Next, literature findings on the responsiveness of the SIP are presented and judged for their validity. The SIP appeared to be able to demonstrate changes in the expected direction and in accordance with changes detected by other instruments. Using data from seven different longitudinal projects in populations with different diagnoses, the responsiveness of both the SIP136 and the SIP68 are subsequently studied and compared. In all populations changes in functional status were indicated by both instruments. In terms of effect sizes the SIP136 and the SIP68 do not differ significantly in their responsiveness. Moreover, changes detected by both SIPs appear to be valid representations of changes in health related functional status.

6.1. Introduction

Traditionally, instruments used to measure health or aspects of health are judged by their level of reliability and validity. When the validity of an instrument is studied, the main question to be answered concerns the accuracy of the instrument as a representation of the concept aimed at. Reliability is concerned with the precision of the assessment. However, when an instrument is to be used in a longitudinal research design, a third aspect becomes important: responsiveness. Responsiveness refers to the level to which an instrument is able to detect changes in the concept measured. In measuring change, just like in measuring health, two criteria can be distinguished: the validity and the reliability of the assessment. In fact, when an instrument is applied in a longitudinal study, the main object for using this instrument is to detect and quantify possible changes in the concept through time, rather than to measure the concept cross-sectionally. If an instrument indicates a change, this should be an accurate representation of the actual change in the concept aimed at. Hence, the ability to accurately assess changes (=responsiveness) in a longitudinal setting, can be viewed as part of the validity of an evaluative instrument. A valid measurement of change presupposes reliability of the assessment, insofar that it should be precise, and not contain a large amount of measurement error.

One of the best known functional status measures is the Sickness Impact Profile (SIP)[1]. The SIP has been studied extensively and was found valid and reliable [2,3]. Relatively few studies have been published in which the responsiveness of the SIP is explicitly studied. Those studies use different methods to assess responsiveness. Most authors suggest further research into the responsiveness of the SIP before final conclusions can be drawn. However, the general tendency in the conclusion of these studies is that the SIP is responsive to changes in functional status [2,4,5,6,7,8]. Recently, a short version of the SIP was developed: the SIP68. This instrument is meant to be a generic alternative for the rather long original SIP [9]. For a clear distinction between both instruments, when SIP136 is mentioned in this paper, reference is being made to the original SIP. As the SIP68 is meant to be an alternative to the SIP136, the responsiveness of the new short instrument will have to be the same as that of the SIP136. In other words, the responsiveness of the SIP136 is an important criterion against which the SIP68 will be judged. Therefore, in this paper the responsiveness of the SIP136 and the SIP68 will be studied and compared. First, both instruments will be briefly introduced. After this, the concept of responsiveness and how to measure it will be addressed. Next, the responsiveness of the SIP136 will be evaluated on the basis of literature findings. Finally, the responsiveness of the Dutch SIP136 and the SIP68 will be assessed and compared, using data from seven longitudinal studies in populations with chronic or lengthy health problems.

6.2. The SIP136 and the SIP68

The Sickness Impact Profile (SIP136) is a generic measure of health related functional status. It consists of 136 items, every item being a statement on behavior. Respondents are asked to check those items that describe their situation on the day they fill out the list, but they do so only when the fact that these items apply to them is related to their health status. The items are grouped into twelve categories, every category covering an aspect of daily functioning. Two dimensions are distinguished: a physical dimension consisting of three categories and a psychosocial dimension composed of four categories. The other categories are not aggregated [1]. Scores for every category, for both dimensions and for the overall instrument are calculated after attaching differential weights to every item. Scores range from 0 to 100 with higher scores representing worse dysfunction. The SIP136 is known to be a valid, reliable and responsive measure of functional status [2,10].

A major disadvantage of the instrument, however, is its length. Therefore, the SIP68 was developed. The SIP68 contains 68 items selected from the original instrument, grouped into six categories. Methods used in this selection procedure have been described extensively elsewhere [9]. Scores on the SIP68 (categories and total scores) are calculated by adding the number of items checked. Evaluation of the SIP68 shows that the instrument measures a concept almost identical to the concept measured by the SIP136. Moreover, the SIP68 meets the same high standards as the original instrument as far as reliability and validity are concerned [9,11,12]. The responsiveness of the SIP68, however, has not yet been explicitly evaluated and compared to the responsiveness of the SIP136.

6.3. Responsiveness

As stated above, the responsiveness of an instrument is the level to which an instrument is sensitive to relevant changes. Hence, two main questions have to be answered before the responsiveness of an instrument can be assessed: (a) against what criterion will the instrument be judged, and (b) what constitutes a relevant change in this criterion that should be registered by the instrument under study?
In the situation where a 'gold standard' exists, the first question is relatively easy to answer. A 'gold standard' is a measure of a concept that is the same as that which the instrument under study aims at, and about which consensus exists concerning its accuracy in representing that concept. Therefore, every change in the standard that is considered relevant should be detected by the instrument under study. The question of what constitutes a relevant change, however, still has to be answered. Although a statistically significant difference between two assessments is conditional in detecting change, not all statistically significant changes will represent a relevant change in the concept. In a situation where no consensus exists concerning the 'gold standard', both questions are difficult to answer. First, a decision has to be made as to what other measure or combination of measures are acceptable as a

criterion. Next, a decision shuold be made as to what changes in this criterion should be registered by the instrument under study. As the criterion in this situation does not measure an identical concept, the question of what constitutes a relevant change is even more difficult to answer than in the situation with a 'gold standard'. For the SIP68, the SIP136 can be considered a natural criterion, as both instruments aim at the same concept, and the SIP136 is generally accepted as valid and reliable. Although the SIP136 is generally said to be responsive [2], no generally accepted procedure exists to judge responsiveness. Therefore, in order to be acceptable as a criterion for the responsiveness of the SIP68, more detailed information is needed on the responsiveness of the SIP136. In fact, in longitudinal studies, the SIP136 repeatedly demonstrated statistically significant changes. Whether these changes are valid representations of relevant changes in health related functional status is hard to determine.
When the SIP68 is compared to the SIP136, however, the second question (what constitutes a relevant change) is relatively easy to answer: analogous with the way criterion validity is assessed in cross-sectional data, a change in SIP68-score should show a high correlation with a SIP136 score change in the same situation. And analogous with the assessment of cross-sectional construct validity, the pattern of relations between SIP68 change scores and changes in related constructs should correspond with the pattern found using SIP136 change scores in relation to measures of these other constructs. Using longitudinal data from different diagnostic groups, the relation between changes demonstrated by both instruments are calculated, as well as relations between the SIP136 and the SIP68 on the one hand and related measures on the other.

6.4. Methodological approaches towards responsiveness

In health status research, roughly two approaches can be distinguished to assess responsiveness: the correlation approach and the responsiveness indexes, developed to compare the responsiveness of instruments.
In correlation studies, the correlation coefficient is calculated between changes in the instrument under study and changes in a well established instrument measuring a related concept. The level of correlation coefficients reveals to what extent changes in one instrument are related to changes in the other. In this situation it is arbitrary what threshold to use above which a correlation coefficient is sufficient to accept the claim for responsiveness. As the SIP68 and the SIP136 are closely related measures, high correlations between change scores on both instruments suggest equal responsiveness.
Several responsiveness indexes have been developed. They all are meant to supply a standardized and dimensionless representation of the change demonstrated by a measure. Table 6.1 presents five different indexes.

Table 6.1. Criteria to judge responsiveness

criteria for responsiveness	author
$(\underline{x}1-\underline{x}2)/s(\underline{x}1)$	Cohen (1977), Kazis (1989)
$(\underline{x}1-\underline{x}2)/s(\underline{x}1-2\underline{x})$	Liang (1990)
min.clin.diff./$s(\underline{x}1-\underline{x}2)$	Guyatt (1987)
$(\underline{x}1-\underline{x}2)/s(\underline{y}1-\underline{y}2)$	Tuley (1991)
(x1-x2)/se(x1-x2)	Jacobson (1992)

$\underline{x}1$: mean assessment at t1
$\underline{x}2$: mean assessment at t2
x1: assessment at t1
x2: assessment at t2
(x1-x2): difference measurement t1 and t2
$(\underline{x}1-\underline{x}2)$: mean difference measurement t1 and t2
$(\underline{y}1-\underline{y}2)$: mean difference measurement t1 and t2 in a stable control group
min.clin.diff.: minimal clinically relevant difference
s: standard deviation
se: standard error of measurement

The five criteria in the above table have a very similar design: the change found or expected is divided by an indicator of the precision of the measurement. Four out of five criteria take the change found on a variable between two measurements as starting point. The first two formulas in table 6.1 are called effect sizes. An effect size is the ratio between the average change found and an indicator of the accuracy of the instrument in question. This denominator might be the standard deviation of the variable at baseline [13,14]. Liang, however, uses an index defined as response mean/response standard deviation (response=difference in score before and after intervention) [7].Effect sizes quantify the amount of change measured. Hence, it seems accurate to use the standard deviation of the change to standardize the amount of change found. The standard deviation at t1 is a cross-sectional figure, and hence less accurate as it does not contain information on the accuracy of the instrument in detecting change over time. In this paper, therefore, when an effect size is calculated, the second type (mean change/ standard deviation of change) will be used. Guyatt suggests an index in which 'the minimal clinically important difference' is the numerator, and the 'between subject variability in within-personal change' is the denominator [15]. A problem in this approach is that what constitutes a 'clinically meaningful change' is very hard to define, or at least, the resulting definition is highly speculative and arbitrary. Comparison of the responsiveness indexes of two instruments reveals to what degree changes in the criterion correspond with changes in the instrument under study as far as the amount of change is concerned. What constitutes a sufficient level of similarity of the effect sizes depends on the expected relation between the instrument under study and the criterion.

Tuley developed a method which determines the statistical significance of the difference in responsiveness between two instruments. He uses a responsiveness index in which the mean change in a treated group is related to the between subject variability of change in a group of stable subjects [16]. Hence, changes in an instrument's score are related to natural variability of scores on that instrument. The expected population value and variance of the index for two different instruments is computed. Next, the difference in expected indexes is calculated. Using data on the variance of the indexes, and on the covariance between the respective scores, a 95% confidence interval of the difference between the two responsiveness indexes is computed. When this interval encloses 0, the difference between the responsiveness of the two instruments is considered not significant [16]. Usually, a researcher is not only interested in statistically significant changes, but in clinically significant improvement or deterioration. To assess the clinical significance of changes detected, Jacobson et al developed the 'reliable change index', an index that assesses the (clinical) relevance of statistically significant changes [17] (fig. 6.1). In contrast with the earlier effect size indexes this is an index at individual level. Jacobson states that if this ratio is greater than 1.96, 'it is unlikely that the post-test score is not reflecting real change' [17]. In other words, if the ratio is less than 1.96, it is not likely that real change has occurred.
All methods mentioned above quantify the change that is demonstrated by an instrument. To be able to judge the validity (or the clinical relevance) of change demonstrated, an external criterion is needed that is acceptable as a standard against which the instrument under study is judged. This external criterion does not exist for instruments like the SIP. Therefore the validity of change can not be proved, but can only be made acceptable.

In this paper, the responsiveness of the SIP136 and the SIP68 will be judged and compared, using Liangs effect size, Tuleys method to compare the instruments, and Jacobsons method to judge the relevance of changes found. First, however, literature findings on the responsiveness of the SIP136 will be presented and discussed.

Figure 6.1. Reliable change index (rc)

$$rc = x_1 - x_2 / S_{diff}$$

S_{diff}= standard error of difference= $\sqrt{2}(SE)^2$
SE= standard error of measurement= $s_1\sqrt{1-r_{xx}}$
s_1= standard deviation of the assessment
r_{xx}= test retest reliability of the measure
(Source: Jacobson et al. 1992)

6.5. Literature findings on the responsiveness of the SIP136

Before the SIP136 is acceptable as a criterion to judge responsiveness, information is needed on the responsiveness of the SIP136 itself. Therefore, findings on the responsiveness of the SIP136 will be presented. Only a few authors report directly on responsiveness. Most studies just use the instrument and when it demonstrates change the assumption is that this indicates actual change in the health related functional status. On the other hand, when it does not demonstrate change the conclusion is that the functional status of the population under study did not change. Only a few studies compare different measures of the same concept for their responsiveness. In this kind of study the question of the criterion against which the different instruments should be judged remains unanswered: they are judged against each other. Usually, the instrument that demonstrates the largest change is considered most responsive. Which instrument demonstrates the most valid change, is a question that remains unanswered. Generally, four types of responsiveness information were found that differ in their conclusive value:
1. Before and after an intervention only the SIP136 is applied, the respective scores are subtracted, and a conclusion is formed as to whether or not the instrument shows change. No criteria are used to support the findings.
2. The SIP136 is used in a study together with instruments measuring other concepts, relevant to this particular study. When different instruments demonstrate change in the same direction, it is considered to be supportive information for the responsiveness of the SIP136. The level of agreement between different instruments is not quantified.
3. The SIP136 is used together with other instruments measuring related concepts. The level of agreement in changes detected by the different instruments is quantified by means of a correlation coefficient. The congruence between instruments, thus is expressed in statistical terms.
4. Finally, several instruments are used and changes found by different instruments are standardized by means of effect sizes. Comparing effect sizes of different instruments yields information on the relative responsiveness of instruments.

Below, data found will be presented according the four categories above. Table 6.2 presents the first type of information: whether or not the SIP136 demonstrated change. Based on the data in table 6.2, it can be concluded that the SIP136 does sometimes show change. When only the SIP136 is used, and no other measure is presented to validate these findings, it is hard to judge to what degree the change registered by the SIP136 mirrors a real change in the health related functional status. In the interpretation of these findings, however, it should be kept in mind that reports on studies that do demonstrate some effect of a treatment might have a bigger chance to be published than studies that do not find any effect. This might lead to the so-called 'publication bias' that might suggest that in the majority of the cases where the SIP136 is used, it does demonstrate a significant effect.

Table 6.2. Findings on changes in scores on the SIP136, not related to other instruments.

author	diagnosis (n)	intervention	time between assessments	conclusion
Ott et al (1983)	myocardial infarction (184)	cardiac rehabilitation	not specified	improvement was egistered
Follick (1985)	chronic low back pain (14 pts. 8 controls)	8 weeks behaviorally oriented rehabilitation	4 months before treatment, immediately after	significant improvement demonstrated
MacKenzie et al (1986)	hospital in-patients with various diagnoses (43)	hospitalization intake, 1, 2, 4,6 weeks	within 24 hrs. after no sign. difference, after discharge scores more sensitive for deterioration than	SIP136 total score SIP136 dimension for improvement
Nielson et al (1990)	elderly undergoing knee arthroplasty (64)	general vs. regional anesthesia	3 months	SIP136 shows 'large' improvements (5% SIPphys; 2 à 5% SIPpsy)
Jacobs et al (1993)	abdominal disorders (980)	initial contact with family physician followed by different treatment courses	1 month, 6 months	SIP showed significant changes

The second and third type of information supplies more convincing information, as these studies compare changes on the SIP136 with changes found with other measures. In table 6.3 the changes demonstrated by different instruments are compared as far as the direction of changes is concerned.
The conclusion from these findings is that in most studies the SIP136 shows changes in the same direction as the other measures do. These findings, however, only provide information concerning whether or not changes in the same direction are detected. Based on these data nothing is to be said about the relative size of the change demonstrated by the SIP136 and other measures.
Slightly more sophisticated information is found in studies that express the relation between changes demonstrated by the SIP136 and related instruments by means of a correlation coefficient. This kind of information is presented in table 6.4.
This table shows that changes registered by the SIP136, generally correlate with changes detected by other measures. Most authors, however, do not specify the level of the correlations found. Hence, it is not possible to differentiate between instruments that show changes that are closely related to changes shown by the SIP136, and instruments that show lower change-score correlations.

Table 6.3. Findings on the direction of changes in the SIP136 score, related to the direction of changes found using other instruments.

author	diagnoses (n)	intervention	time between assessments	other measures	conclusion
Turner (1982)	low back pain (36)	waiting list vs. relaxation or cognitive therapy (5 wk.program)	baseline, 5 weeks, 1 month	SIP-proxy, Beck Depression Inventory v.a.s.-pain	SIP136 and other measures improved significantly in therapy groups, not in waiting list
Deyo (1983)	low back pain (80)	treatment in walk-in clinic	3 weeks	clinical measures/ judgements, Likert scale pain, self-assessed change	pts. judged 'worse' had worse SIP136 score 'same' or 'better' both improved on SIP136
Kubo et al (1988)	congestive heart failure (17)	opc-8212 (drug) treatment	4 weeks	Minnesota living with heart failure questionnaire (MLHFQ), exercise duration, peak oxygen consumption	SIP136 and MLHFQ showed sign. change, other measures did not
Berwick et al (1989)	low back pain (222)	back school with or without 'compliance package'	baseline, 3, 6, 12, 18 months after enrollment	v.a.s.-pain level	no effect demonstrated on SIP136 or other measures
Hyde (1989)	pregnant women with morning sickness (16)	acupressure wristband vs. no therapy	5 days	Multiple Affect Adjective Checklist, extend of nausea	SIP136 showed sign. changes just like the other measures
Tandon et al (1989)	congestive heart failure (111)	standard vs. placebo therapy	12 weeks	self rating scale, Quality of Life Index (QLI), Quality of Wellbeing scale (QWB)	SIP136 did show some change, QWB did not

Finally, some authors calculate and compare effect sizes found using the SIP136 and related instruments. Effect sizes enable comparison of the size of changes detected by different instruments. Liang compared effect sizes of 5 global health measures (SIP136, Arthritis Impact Measurement Scale (AIMS), Functional Status Index (FSI), Health Assessment Questionnaire (HAQ), Index of Well Being (IWB)) in a population with end stage arthritis that undergo hip or knee arthroplasty [7]. It is concluded that the SIP136, the HAQ and the IWB are suitable to register change (effect sizes respectively: 1.11, 1, 0.88) Turner found significant changes in a study in chronic low back pain patients receiving behavioral exercise [31]. In this study the SIP136 did not demonstrate significant changes, and its effect size is smaller than that

of the Pain Behavior Checklist (PBC), the McGill Pain Questionnaire (MPQ), observer rated pain and measures rated by spouses. Jacobs found effect sizes ranging from 1.01 to 0.03 in a population of abdominal complaints of different nature [20]. According to Cohen's rule of thumb, an effect size of 0.20 is considered small, 0.50 moderate and 0.80 or more is judged as large [13]. Judged by this classification, the SIP136 demonstrates small as well as large changes.
A general problem in the discussion on change scores and the detection of change, is the phenomenon of 'regression to the mean', and the so called 'ceiling- or floor effects'. Supposing a certain amount of chance- or error-variance in every assessment, respondents that initially obtain a very high score (at the ceiling of the scale) are more likely to obtain a lower score in the retest because they can not go higher (through the ceiling). The reverse is also true: respondents obtaining a very low score (at the floor of the scale) are more likely to obtain an increased score at the retest because they can not go 'through the floor'.
As far as the SIP136 is concerned, due to the fact that the instrument has a generic character and is very broad in its scope, it is almost impossible to obtain the maximum score. On the contrary, usually scores are skewed towards the lower end of the scale. This is caused by the fact that a certain type of ill-health is accompanied by impacts in a limited range of functional categories. Hence, it might be expected that the SIP136 is more likely to show deteriorations (higher scores represent worse functional status) than to detect improvements. Only MacKenzie et al found that the SIP136 dimension scores are more sensitive for deteriorations than for improvement [18]. Other authors do not mention this 'skewed responsiveness' in connection with the SIP136.

From the above it appears that in many different types of populations and research settings, the SIP136 demonstrates changes in the expected direction and these changes are (generally) in accordance with changes detected by other measures. However, studies in which this type of result is found, are generally not designed to study the responsiveness of the instrument. Hence, the above findings can only be interpreted as indications of responsiveness. However, the preliminary conclusion based on these findings is that the SIP136 is able to show changes and that these changes appear to correspond with changes found using other measures. The validity of these changes and the validity of the size of change registered by the SIP136, still has to be studied. This, however, has to be said about all instruments that measure abstract concepts like 'general health' and 'functional status'. Bowling, in her review of health status measurements formulates a general conclusion that is in accordance with the above findings: the SIP136 'is sensitive to change', and 'is valuable (...) for measuring the effects of non-curative interventions' [2]. In this paper this claim is not denied, but, as stated above, there is not (yet) sufficient evidence to firmly support this statement. In the following section responsiveness of the SIP136 and the SIP68 will be systematically studied and compared using empirical data.

6.6. Studying the responsiveness of the SIP136 and SIP68

6.6.1. Data

To study the responsiveness of the SIP136 and the SIP68, data will be used from seven longitudinal studies. In these studies the validated Dutch SIP136 [32] and a number of self-assessment Likert-scales assessing aspects of health, were administered. These Likert scales will serve as a criterion against which both SIP versions will be judged. Table 6.5 (appendix) presents an overview of the available populations. A horizontal line separates the studies. The first project 'after the rehabilitation center' is a study into the course of functioning after discharge from (clinical) rehabilitation [33]. In this study respondents were interviewed at admission to the rehabilitation center, at discharge, three months after discharge and two years after admission. Diagnoses represented in this population are rheumatoid arthritis, spinal cord lesion and ankylosing spondylitis. The spinal cord injury population did not fill out the SIP136 at admission, therefore change scores are calculated using data from the third and fourth measurement moment. Change scores from both other diagnostic groups are calculated using data from measurements before and just after admission.
The second study is a project that evaluates the efficacy of back-schools [34]. The next study is a project comparing the effect of group physical therapy and exercise at home in a population of ankylosing spondylitis patients [35]. The fourth population stems from a study that compares the effectiveness of physical therapy and manual therapy in a population with back and neck complaints [36]. The fifth group stems from a study into the quality of life of cancer patients in the first year after diagnosis [37]. The stroke population contains patients from an academic hospital who were followed during half a year [38]. Functional status assessment took place at six months and 12 months after the cerebral accident. Finally a population was used that stems from a study into personal networks of chronic patients [39]. Half of this population was diagnosed as having Crohn's disease, the other half was known to have ankylosing spondylitis. Both groups received standard therapy for their respective ailment. Twice, half a year elapsed between the assessments.

SIP136 scores are calculated following the official scoring procedure, where differential weights are attached to the items, and the percentage of the total possible score is calculated (theoretical range 0% to 100%). The SIP68 total score is calculated by adding the number of items checked (theoretical range 0 to 68). Hence it is not possible to compare the scores directly. Table 6.5 indicates that total scores on the SIP136 in almost all populations is around one point higher that the total score on the SIP68. Considering the difference in length and scoring method between both instruments it is surprising that the difference is that small. However, in previous publications we also found that scores on the SIP68 were closely related to scores on the SIP136 [9,11]. As the populations stem from different studies, not all criterion variables were used in every population. All criteria mentioned are self-assessments of aspects of health on five-point Likert scales. The changes found using these scales can serve as a criterion against which the responsiveness of the SIP136 and SIP68

can be judged. In every population change scores were calculated for both SIP versions and for the available criterion measures. Table 6.5 presents changes at a total score level for both SIP versions. Change scores were calculated by subtracting the score of two successive administrations (score at t1 minus score at t2). As a higher SIP score indicates a worse functional status, a positive change score indicates an improvement in the functional status. In table 6.6 (appendix) effect sizes are presented for statistically significant changes at the level of category-, dimension- and total score. From this table it appears that all populations show a statistically significant change in at least one category or dimension. All changes demonstrated are positive, indicating an improved functional status. In all situations where the SIP136 or one of its dimensions show a change, the SIP68 total score changes in the same direction and the extent of the changes is approximately equal.

In table 6.5 significant changes found using both instruments are presented for total scores. Table 6.6 (appendix) presents the effect sizes of significant changes found using the SIP136 and the SIP68 at the level of categories and dimensions. All effect sizes (except one category es) are positive, indicating improvement of functional status. Tables 6.5 and 6.6 show that the largest changes are found in populations that obtain the highest initial score. This finding is congruent with the expected 'floor effect': a population that initially obtains a high score on the SIP136 or SIP68 (=severely affected functional status), is able to gain more by treatment than a population that is only mildly affected initially. Comparing changes registered by both the SIP136 and the SIP68, it can be concluded that when the SIP136 score changes, the SIP68 score also changes. Differences found with both instruments point in the same direction. As far as the magnitude of differences (the effect sizes) is concerned, however, SIP136 changes consequently appear to be larger than changes found using the SIP68.

6.6.2. Correlation approach

Mean change score comparison is a rather rough way of comparing the responsiveness of two instruments. The correlation coefficient between two change scores, supplies more insight into the relation of two instruments as far as the responsiveness is concerned. Table 6.7 presents the correlation coefficients between significant changes found by the SIP136 and the SIP68. The second column of this table indicates a high to very high correlation between changes in the total score of both instruments, indicating a high level of agreement between changes detected by both instruments.

Correlations between changes in the SIP68 total score and the SIP136 dimension scores are lower. This difference in agreement will be due to the fact that the SIP68 aims at the total concept of health related functional status, while both SIP136 dimension scores are aspects of this overall concept. When the SIP136 is chosen as criterion against which the SIP68 is validated, these findings support the criterion validity of the changes detected by both instruments.

Table 6.7. Pearson's correlations between significant changes registered by SIP136, its dimensions and the SIP68

p< .01

diagnosis	SIP136 - SIP68	SIP68 - phys.dim	SIP68 -psy.dim.
rheumatoid arthritis	.90	.85	.75
ankylosing spondylitis	.94*	.77	.80
spinal cord injury	.92	.75	.69
chronic back pain			
ankylosing spondylitis	.89	.59	.76
back and neck complaints	.85	.85	.65
cancer	.90	.76	.72
stroke			
Crohn's disease	.94	.77	.81
ankylosing spondylitis			

* p < .05

A next step would be to compare changes registered by both instruments in relation to changes found using some measures of related concepts. When a functional status measure is compared to some physiological measure, a very high correlation between both measures is not desirable because the concept of functional status is not identical to the concept measured by physiological measures. On the other hand, a certain level of correlation is expected, because on theoretical grounds one does expect a certain congruence between these two aspect of health. Table 6.8 (appendix) presents correlation coefficients between significant changes detected by both SIP versions and the available self-assessment Likert scales. In more than half the number of SIP-criterion combinations, no significant correlation was found. This might be due to the fact that changes demonstrated by the SIP and the criteria are expressed on very different scales. A SIP136 score potentially ranges from 0 to 100%, a SIP68 score potentially ranges from 0 to 68. The largest possible difference on a five point scale is 4. Hence, it can be expected that if any change occurs, both SIPs will be more subtle in its registration than five point Likert scales. Based on this difference in scales and the relatively small changes registered by both SIPs, it is not to be expected that many significant and high correlations will be found. The correlations that were found all pointed in the expected direction: a positive change in SIP-score coincides with a positive change on the Likert scales. Highest correlations are found between both SIPs and self-assessment of health and physical functioning, slightly lower coefficients are found with self-assessment of hindrance, psychological functioning and severity. For every Likert scale except for the self assessed level of pain, significant correlations are found between change scores. Functional status is chosen as an operationalisation of health in general, hence a high correlation between these two is expected. As functional status in the SIPs is expressed in terms of behavior, a close relation is to be expected between

the SIP-change scores and changes in self-assessment of physical functioning. The relation between functional status and psychological functioning is expected to be less direct. Hindrance and severity are very general, non-specific and abstract concepts that are expected to have a more diffuse relation to functional status expressed in behavioral terms. The relation between the experience of a certain level of pain and the consequences of this pain on someone's behavior is very complex, and individually different. Thus the fact that no significant correlation was found between self assessed pain and functional status, is not surprising.
In general, when a significant correlation is found between SIP136 and a self-assessment, a correlation of similar height is found between the SIP68 and this scale. For both SIPs it can be concluded that this pattern confirms the construct validity of the changes measured.

6.6.3. The sizes of changes measured

For every population in table 6.5 the reliable change index was calculated for the total scores on both the SIP136 and the SIP68 following Jacobson (17). None of the statistically significant differences in table 6.6 yield an index that is greater than 1.96. Hence, if the rather strict norm of Jacobson is applied, it appears that in none of the studies a clinically significant change was detected. Although no spectacular changes are found in the populations in the present study, still, the size of changes found using both SIP versions can be compared by means of effect sizes. Table 6.6 (appendix) presents the effect sizes registered by both instruments. Effect sizes are calculated for every significant difference in category, dimension or total score found between two assessments. According to the rule of thumb suggested by Cohen (see above) all effect sizes found are small or at most moderate. In all populations except the spinal cord injured population, the SIP136 demonstrates a slightly larger effect size than the SIP68. These differences, however, are small.
Using the method Tuley developed to compare the responsiveness of two instruments and data from two of our populations (the stroke population not treated between the assessments, and the population containing Rheumatoid Arthritis, Ankylosing Spondylitis and Spinal Cord Injury as treated), we assessed the statistical significance of the difference in responsiveness between both instruments. We found that responsiveness indexes for SIP136 and SIP68 respectively were 0.64 and 0.62. The 95% confidence interval for the difference between these outcomes ranges from -0.118 to 0.112, with a mean of -0.0033. As this interval encloses zero, the conclusion is that there is no significant difference between the change registered by the SIP136 or the SIP68 respectively.

6.7. Discussion and conclusion

The subject of this paper is responsiveness. After a general introduction in the subject and an inventory of possible methodological approaches, responsiveness of the SIP136 was evaluated using literature findings. Next, using empirical data from several sources, the responsiveness of the SIP136 and the SIP68 (a short version) were studied and compared. It was argued that as the SIP68 is meant to be an alternative for the SIP136, this last instrument can serve as a criterion against which the first is judged. Literature findings indicate that the SIP136 is probably sensitive to relevant changes in the health related functional status. Conclusive and firm evidence for the responsiveness of the SIP136 was not found. Analysis of data from seven longitudinal studies in which the Dutch SIP was used showed that in all seven studies significant but relatively small differences were registered by both SIP versions. A high correlation was found between changes detected by both SIP versions, suggesting that both instruments are approximately equally responsive. As far as direction and level of correlation coefficients are concerned, the pattern of correlations between change scores of both SIPs and measures of several aspects of health (self-assessments on Likert scales), is congruent with theoretical expectations. Also, the magnitude of changes detected was evaluated. Using the rather strict interpretation of the 'reliable change index' of Jacobson, none of the changes found by either SIP version appeared to be clinically significant. When effect sizes of both instruments were studied, it was found that there was no statistically significant difference between effect sizes of both SIPs.
Based on these findings it can be concluded that both instruments are equally accurate in detecting changes in the health related functional status. Moreover, it is highly probable that changes detected by the SIPs are valid representations of changes in the health related functional status as they relate to changes in measures of related concepts in the expected way.

The fact that only small, if any, changes were found might be due to the fact that the populations used all had chronic or lengthy conditions and actually did not change much between the measurement moments. In that case both SIP versions adequately and validly detected changes. As it is not to be expected that an instrument that accurately detects relatively small changes would be less suitable to detect large changes, both the SIP136 and the SIP68 can be considered sufficiently responsive. The generic character of both SIPs also might be due to the fact that only small changes were demonstrated. A generic measure is to be used in populations with all types and severities of bad health. Therefore, only rather general possible functional consequences of bad health are checked by the instruments. Subtle functional nuances, characteristic for specific health problems, hence, might be 'overlooked' and the actual change is underestimated. Therefore, before definitive conclusions can be drawn concerning the responsiveness of the SIP68 and the SIP136, responsiveness of these instruments should be judged in relation to changes demonstrated by disease-specific measures of functional status.

A more basic consideration is that by its nature, the concept at which the SIP aims (health related functional status), is relatively stable and does not change easily. It is well-known that there is no uni-dimensional and generalizable relation between a given state of health in bio-physical terms and connected behavioral changes. However firm a medical diagnosis may be, hardly any officially codified disturbance of health (diagnosis) is connected with a clear-cut pattern of behavioral changes. All health care professionals know that the behavioral consequences of a given diagnoses may vary strongly between individuals. This might be caused by the fact that behavioral patterns are persistent and are only marginally influenced by most health changes.

A person's behavior is not only influenced by his bio-physiologic health status, but also by his social situation. The kind of work he is used to doing, his coping strategies, restrictions and possibilities in the domestic situation, etcetera. Especially respondents in a chronically unhealthy situation will have developed a more or less stable pattern of behavior that fits their situation. Hence, only intensive treatment or large and obvious changes in the disease might lead to a statistically significant and clinically relevant change in the SIP-score. In this study the largest changes are found in populations receiving intensive treatment (clinical rehabilitation) and in a population suffering from a far reaching disease (cancer). Clinical rehabilitation is an intensive kind of multi-disciplinary treatment, directed at minimalizing the (functional) consequences of bad health, not at removing the cause of disease. Often, spouses or other informal care takers are involved in the rehabilitation process. Given these characteristics, it is to be expected that in this population new negotiations will start on the true functional consequences of a given state of health. The cancer population was followed for one year after the initial diagnosis. Given the character of this disease, it is to be expected that during the first year part of the population present at t1 will have died within this year. The surviving part will have lived through a tempestuous year as far as their health and percepted functional possibilities are concerned. Hence it is to be expected that this population also shows relatively large changes on the SIP. The other populations receive treatment that is less intrusive, or they are in a more stable period of their disease (eg. stroke patients at 6 months and one year after the stroke, receiving no therapy). Hence, given the stable character of the SIP concept, it is to be expected that in these populations only minor changes in SIP scores would be found.
These considerations about the nature of the concept measured by the SIP (both SIP136 and SIP68), lead to the conclusion that application of these instruments in longitudinal research might be more suited when large changes are expected and when the period between measurements offers the possibility of redefinition of the level of health related functional status. However, as stated before this is a theoretical point. Further research on this aspect is needed before the impact of this topic on the responsiveness of the SIP and related instruments can be described in more detail.

Table 6.4. Data found in literature on correlations between SIP136 change-scores and changes-scores on other measures.

author	diagnosis (n)	intervention	time between assessments	other measures	conclusion
Rockey et al (1980)	hyperthyroid pts. (14)	medical treatment	4 to 8 weeks intervals until pt. is euthyroid	serum T4 level	strong correlation SIP136-T4 level SIP136 approaches 0 when pts. euthyroid
Deyo et al (1984)	rheumatoid arthritis (79)			self rating, American Rheumatism Association scale, self-assessment of change, agreed upon change-score	all scales relatively insensitive: SIP changes small. Only SIPphys. en pt.self rating correlate sign. with clinically estimated change
Deyo et al (1986)	acute low back pain (120)			Roland scale, self/prof. rating of improvement, spine flexion, straight leg raising, resumption full activities	corr. SIP with self-assessment/spine flexion/ leg raising: low ± .30 (dimensions less)
Canadian Erythropoietin study group (1990)	anaemic patients (118)	hemodialysis and erythropoietin or placebo	2 resp. 6 months	kidney disease questionnaire, time trade off, 6 minute walk, modified Naughton stress test	correl. found hemoglobin change and SIP tot,-phys,-psy change
Caradoc-Davies et al (1990)	injured people attending multi disciplinary rehabilitation center (32)	multi-disciplinary rehabilitation (8 weeks)	8 weeks	fitness for work (ACC classif.), occupational handicap (ICIDH), program outcome (6-points), client perception of benefit	benefit correlated sign. with SIP change

Table 6.5. Populations used in this study

diagnosis	*n*	*intervention*	*time span*[1]	*criteria*[2]	*SIP136 t1(sd)*	*change 136*(sd)*	*SIP68 t1(sd)*	*change 68*(sd)*
rheumatoid arthritis	34	clinical rehabilitation	duration of clinical admission	1,2,3,4,5	26.3 (13)	4.4 (8.0)	25.2 (8.1)	2.9 (6.0)
ankylosing spondylitis	29	clinical rehabilitation	duration of clinical admission	1,2,3,4,5	11.7 (8.9)	3.9 (5.3)	10.6 (8.7)	3.5 (5.1)
spinal cord injury	19	clinical rehabilitation	approximately 15 months	1,2,3,4,5	20.7 (8.6)	4.4 (6.0)	19.2 (6.7)	4.8 (6.4)
chronic back pain	76	back school	6 months	6	5.9 (7.5)	ns	4.6 (6.5)	ns
ankylosing spondylitis	133	physical therapy	9 months	1	4.8 (4.7)	1.1 (3.5)	4.0 (4.5)	.88 (3.1)
back and neck complaints	228	physical or manual therapy	6 weeks	5	4.8 (4.6)	1.9 (4.1)	2.3 (3.4)	.98 (3.1)
cancer	60	standard treatment	1 year		9.5 (8.5)	3.1 (5.7)	8.3 (7.9)	2.5 (5.3)
stroke	51	no treatment	6 months	1,2,3,4,5	15.5 (11.0)	ns	14.9 (9.6)	ns
Crohn's disease	37	standard treatment	1 year	1,2,3,5	5.6 (6.6)	2.1 (5.1)	4.4 (5.6)	1.6 (4.8)
ankylosing spondylitis	40	standard treatment	1 year	1,2,3,5	9.2 (7.5)	ns	8.0 (7.2)	ns

sources: 27,28,29,30,31,32,33.
* only significant differences between t1 and t2 are printed
sd: standard deviation
ns: not significant
[1] time span between assessment used to calculate change scores
[2] 1 self-assessment of health on a Likert scale
2 self-assessment of hindrance on a Likert scale
3 self-assessment of physical functioning on a Likert scale
4 self-assessment of psychological functioning on a Likert scale
5 self-assessment of severity of illness on a Likert scale
6 self-assessment of pain on a Likert scale

Table 6.6. Effect sizes based on significant differences measured using SIP136 and SIP68

effect size= (x1-x2)/s(x1-x2)

	SR	EB	BcM	HH	Mob	SI	Amb	AI	Comm	RP	E	phys	psy	136 tot	SA	MC	PAC	SB	ES	MR	68 tot
1	.54					.35		.38			.41	.37	.46	**.55**				.39			**.48**
2	.54		.86		.37	.40				.66		.83		**.73**	.73	.40		.68	.52	.45	**.68**
3						.68				.57		.51	.54	**.73**				.65			**.75**
4	.25				.26											.28					
5	.16	.17				.22			.20	.37	.21		.25	**.31**				.28			**.28**
6	.39	.51	.45	.12		.16	.17			.26		.39	.32	**.46**	.25	.34		.28	.20		**.31**
7	.52			.45	.43	.31						.35	.34	**.54**				.63		.27	**.47**
8			.29																		
9			.42	.44		.35							.36	**.41**	.43				.43	.44	**.33**
10							-.36														

1: reumatoid arthritis
2: ankylosing spondilitis
3: spinal cord injury
4: chronic back pain
5: ankylosing spondilitis
6: back and neck complaints
7: cancer
8: stroke
9: Crohn's disease
10: ankylosing spondilitis

SR: sleep & rest
EB: emotional behavior
BcM: bodycare & movement
HH: household management
Mob: mobility
SI: social interaction
Amb: ambulation
AI: alertness & intellectual functioning
Comm: communication
RP: recreation & pastimes
E: eating

phys: physical dimension
psy: psychosocial dimension
136tot: SIP136 total score
SA: somatic autonomy
MC: mobility control
PAC: psychic autonomy & communication
SB: social behavior
ES: emotional stability
MR: mobility range
68 tot: SIP68 total score

Table 6.8. Pearsons correlation coefficient between change scores on SIP136-SIP68 and criterion variables

p<.01; *:p<.05

diagnosis	Health		hindrance		phys. funct.		psy.funct.		severity		pain	
	SIP136	SIP68	SIP136	SIP68	SIP136	SIP68	SIP136	SIP68	Sip136	SIP68	Sip136	SIP68
rheumatoid arthritis							.43*	.43*			--	--
ankylosing spondylitis											--	--
spinal cord injury	.62	.62					.52*		--	--	--	--
chronic back pain	--	--	--	--	--	--	--	--	--	--		
ankylosing spondylitis			--	--	--	--	--	--	--	--	--	--
back and neck complaints	--	--	--	--	--	--	--	--	.27	.28	--	--
cancer	--	--	--	--	--	--	--	--	--	--	--	--
stroke			.34*	.37	.38	.40			.30*	.49	--	--
Crohn's disease	.61	.57	.37		.69	.64	--	--			--	--
ankylosing spondylitis	.39*	.38*	.56	.42	.43		--	--			--	--

When nothing is printed in a cell no significant correlation was found.
-- criterion measure not available.

Literature

1. Bergner M, RA Bobbitt, WB Carter, BS Gilson. The Sickness Impact Profile: Development and final revision of a health status measure. Med Care 1981, 19: 787-805.
2. Bowling A. Measuring health, a review of quality of life measurement scales, Buckingham, Open University Press, 1991.
3. de Bruin AF, LP de Witte, FCJ Stevens, JPM Diederiks. Sickness Impact Profile: the state of the art of a generic functional status measure. Soc. Sci. Med. 1992, 35: 1003-1014.
4. Ott CR, ES Sivarajan, KM Newton, MJ Almes, RA Bruce, M Bergner, B Gilson. A controlled randomized study of early cardiac rehabilitation: The Sickness Impact Profile as an assessment tool. Heart & Lung march 1983, vol 12, no.2.
5. Hart GL, RW Evans. The functional status of ESRD patients as measured by the Sickness Impact Profile. J. Chron. Dis. 1987, 40: 117S-130S.
6. Follick MJ, TW Smith, DK Ahern. The Sickness Impact profile: a Global Measure of Disability in Chronic Low Back Pain. Pain 1985, 21: 67-76.
7. Liang MH, AH Fossel, MG Larson. Comparisons of five health status instruments for orthopedic evaluation. Med. Care 1990, 28: 632-642.
8. Deyo RA, RM Centor. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J. Chron. Dis. 1986, 39: 897-906.
9. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen. The development of a short generic version of the Sickness Impact Profile. J. Clin. Epid. 1994, vol. 47, no.4: 407-418.
10. McDowell I, C Newell. Measuring health: a guide to rating scales and questionnaires, New York, Oxfort University Press, 1987.
11. de Bruin AF, M Buys, LP de Witte, JPM Diederiks. The Sickness Impact Profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. J. Clin. Epid. 1994, vol. 47, no.8: 863-871.
12. Post MWM, AF de Bruin, LP de Witte, G Schrijvers. The SIP68: A measure of health-related functional status in rehabilitation medicine. submitted.
13. Cohen J: Statistical power analysis for the behavioral sciences, New York: Academic Press, 1977.
14. Kazis LE, JJ Anderson, RF Meenan. Effect sizes for interpreting changes in health status. Med. Care 1989, 27: Supplement.
15. Guyatt G, S Walter, G Norman. Measuring change overtime: assessing the usefulness of evaluative instruments. J. Chron. Dis. 1987, 40: 171-178.
16. Tuley MR, CD Mulrow, CA McMahan. Estimating and testing an index of responsiveness and the relationship to power. J. Clin. Epid. 1991, 44.
17. Jacobson NS, P Truax. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. in: Kazdin AE. Methodological issues & strategies in clinical research. American Psychological Association, Washington DC, 1992.
18. MacKenzie CR, ME Charlson, D DiGioa, K Kelley. Can the Sickness Impact Profile measure change? an example of scale assessment. J. Chron. Dis. 1986, Vol. 39, no.6: 429-438.
19. Nielson WR, AW Gelb, JE Casey, FJ Penny, RN Merchant, PH Manninen. Long term cognitive and social sequelae of general versus regional anesthesia during arthroplasty in the elderly. Anesthesiology 1990, 73: 1103-1109.
20. Jacobs HM. Health status measurement in family medicine research. Thesis, University of Utrecht, 1993.
21. Turner JA. Comparison of group progressive relaxation training and cognitive-behavioral group therapy for chronic low back pain. Journal of Consulting and Clinical Psychology 1982, vol. 50, no.5: 757-765.
22. Deyo RA, AK Diehl. Measuring physical and psychosocial function in patients with low back pain. Spine 1983, vol. 8, no.8: 635-642.
23. Kubo SH, TS Rector, JE Strobeck, JN Cohn JN. OPC-8212 in the treatment of congestive heart failure: results of a pilot study. Cardiovascular Drugs and Therapy 1988; 2: 653-660.
24. Berwick DM, S Budman, M Feldstein. No clinical effect of back schools in an HMO, a randomized prospective trial. Spine 1989; vol 14, no.3: pp 338-344.

25. Hyde E. Acupressure therapy for morning sickness, a controlled clinical trial. Journal of Nurse-Midwifery 1989; vol 34, no 4.
26. Tandon PK, H Stander, RP Schwarz. Analysis of quality of life data from a randomized, placebo-controlled heart failure trial. J Clin Epid 1989; Vol 42, no 10: pp.955-962.
27. de Witte LP. After the rehabilitation centre a study into the course of functioning after discharge from rehabilitation (dissertation) Amsterdam/Lisse. Zwets en Zeitlinger. 1992.
28. Janssen M. Personal networks of chronic patients. PhD. thesis, Maastricht: University of Limburg, 1992.
29. Hidding A, Sj van der Linden, M Boers, X Gielen, A Kester, LP de Witte, B Dijkmans, J Moonenburgh. Is group physical therapy superior to individual therapy in ankylosing spondylitis. A randomized controlled trial. Arthritis Care and Research, submitted.
30. Visser-Meily A, MGPM Geerts. Beloop en beperkingen van CVA-patiënten gedurende het eerste jaar na het CVA (The course and limitations of functioning of cerebrovasculair accident patients during the first year after the cerebrovascular accident). presentation at the 'fall symposium VRA'. Hoensbroek: 30-10-1992.
31. Courtens AM. Kenmerken van zorg en kwaliteit van leven bij patiënten met kanker (Characteristics of Care and Quality of Life in cancer patients), PhD. thesis, Maastricht: University of Limburg, 1993.
32. Keijzers JFEM. The efficacy of backschools: empirical evidence and its impact on health care practice. PhD. thesis, Maastricht: University of Limburg 1991.
33. Koes BW. Efficacy of manual therapy and physiotherapy for back and neck complaints. PhD. thesis, Maastricht: University of Limburg, 1992.

Chapter 7

Analyses on the items dropped during the SIP68 selection procedure

A.F de Bruin[1], J.P.M. Diederiks[1,2], L.P. de Witte[1,2], F.C.J. Stevens[1], H. Philipsen[1].

[1] University of Limburg, Department of Medical Sociology, Maastricht, The Netherlands
[2] IRV, Hoensbroek, The Netherlands

7.1. Introduction

In the SIP project, a short version of the Sickness Impact Profile [1] was developed [2]. This short instrument contains items that were selected from the original 136 SIP-items. Original items were deleted for three successive reasons:
1. items in the category 'work' were not considered relevant for a large part of the potential respondents, and not suitable for the Dutch situation;
2. a second number of items was deleted because they did not differentiate within the populations since the answering pattern for these items was very skewed;
3. finally, items were deleted because of their relatively low factor loadings in the solution of a principal components analyses.

The first two reasons differentiate relevant items from less relevant items: the category 'work' does not supply usefull information in the Dutch situation, and skewed items do not contribute essential information on the differences in functional status within a population. The third reason, however, (the low factor loadings) is of a more arbitrary nature. A low factor loading indicates little relevance for the component measured by a particular factor. This does not automatically imply that the item in question is not a relevant indicator of functional status. Although it was argued during the development of the SIP68 that no essential information from the original SIP was lost in the selection procedure [2], in this appendix the psychometric characteristics and theoretical relevance of the 40 items dropped because of low factor loadings were studied using the same populations that were used to develop the SIP68. The question answered in this appendix is whether the 40 items dropped because of low factor loadings (rest items) contain information that is essential for the measurement of health-related functional status.

7.2. Comparison of SIP136, SIP68 and 'SIPREST'

Both the SIP68 and the rest items contain items that stem from the original SIP and that are considered to contribute to the differentiation of the level of health-related functional status (no skewed answering pattern). Therefore, it is to be expected that both groups of items in some way represent the concept measured by the original instrument. The question, however, is whether the SIP68 contains all relevant aspects measured by the SIP136. It might be that some aspect of functional status was excluded from the SIP68 due to choices made during the selection procedure. To investigate whether the rest items contain aspects of health-related functional status that are not incorporated in the SIP68, a principal components analysis was performed on the 40 rest items. This analysis might reveal subsets of items that represent aspects of the total concept measured.
Without restricting the number of factors to be extracted, 13 factors appeared to be present. The content of these factors could not be interpreted as 13 seperate aspects of functional status. It was hypothesized that the rest items are 'the other half' of the relevant SIP136 items, and hence it could be expected that they represent

the same concept as the SIP68 does, using other items. Cronbach's α of the list of 40 rest items was .81, indicating an sufficient level of internal consistency to assume that the total rest list, just like the SIP68, measure one total concept. Consequently it was decided to restrict the number of factors to be extracted to 6, the same number as in the SIP68. Data on the resulting factor solution is presented in tables 7.1 and 7.2 (appendix). On the basis of the eigenvalues of the factor and the amount of variance explained by the factors (table 7.2), it can be said that the factor solution of the rest items is slightly inferior to that of the SIP68, as both eigenvalues and amount of explained variance of the rest item factors are generally less high than in the SIP68 factor solution. Table 7.2A (appendix) presents data on the internal consistency of the six factors. All factors have a Chronbach's α of less than 0.70, indicating a relatively low level of internal consitency. Inspection of the content of the factors confirmed this finding, since only items loading on factor 4 might be interpreted as representing a more or less homogenous aspect of health or functional status. The contents of items loading on each of the other five factors are too heterogeneous to represent one functional status aspect. The fourth factor, however, can be interpreted as a representation of 'social isolation'. The five items that load on this factors are concerned with staying alone, showing less affection or interest, and reduction of visits. It was argued that 'social isolation' might be a theoretically relevant aspect of health-related functional status. Social isolation is not represented in the SIP68 as a separate category. Therefore, adding the category 'social isolation' would improve the theoretical validity of the SIP68 and the empirical coverage of the concept of health-related functional status of the instrument. Hence, the five 'social isolation' items were added to the 68 SIP68 items and the resulting 'SIP73' was studied as an alternative short SIP136 version. Table 7.3 presents the scores found using the different SIP-versions.

Table 7.3 shows that the raw scores on the SIP68, the 'SIP73' and the SIP136 do not differ much. The score on the 40 item 'SIPrest' list does differ from the results of the other instruments. However, direct comparison of these scores does not render relevant information as the lists differ greatly in theoretical range.

Table 7.3. Total scores on the 'SIPREST', SIP68, 'SIP73' and SIP136

	mean score (st.dev)		median	theoretical range
'SIPrest' (40 rest items)	6.4	(4.9)	5.0	0 - 40
SIP68	11.5	(9.6)	9.0	0 - 68
'SIP73' (SIP68 + 5 SI items)	12.2	(10.3)	10.0	0 - 73
SIP136 (percentage score)	11.3	(9.5)	8.7	0 - 100
SIP136 (dichotomous score)	18.5	(14.3)	16.0	0 - 136

Table 7.4. Correlation coefficient of 'social isolation' with SIP68, and its categories

	SIP68	SA	MC	PAC	SB	ES	MR
social isolation	.57	.24	.35	.40	.55	.49	.46

The correlation coefficient between the three sets of scores would indicate the level to which information supplied by the instruments is related. As the 40 'SIPrest' items are not supposed to represent a relevant construct, only the SIP68 and the 'SIP73' are compared to the SIP136, and to each other. The correlation coefficient between the SIP136 and SIP68 is .96; correlation between the SIP136 and 'SIP73' is also 0.96 (Pearsons' r, $p< .00$). The correlation between the SIP68 and 'SIP73' appeared to be extremely high: 0.99 (Pearsons'r, $p< .00$). From these findings it can be concluded that scores on the SIP68 and 'SIP73' are very closely related. It might even be stated that both lists measure an identical concept. Adding the 'social isolation' items to the SIP68, apparently does not essentially change or improve the SIP68 as a measure of health-related functional status. This is confirmed by the fact that the relation to the SIP136 is not influenced by adding SI. However, since social isolation could be considered a relevant aspect of health-related functional status, the relation between this category and the six SIP68 categories is studied. Table 7.4 presents correlation coefficients between scores on Social Isolation and the SIP68 categories and total scores.
Table 7.4 shows that Social Isolation is most closely related to the SIP68 total score (r=.57) and the SIP68 categories Social Behavior (r=.55), Emotional Stability (r=.49) and Mobility Range (r=.46). It might be that behavior that leads to social isolation is represented in the more general SIP68 categories. Moreover, it might be hypothesized that behavior described in SB, ES and MR results in social isolation. This hypothesis is supported by the results of a regression analysis in which the score on SI is predicted on the basis scores of the SIP68 categories. In a stepwise regression procedure the categories SB, ES, MR and PAC are entered in the regression formula. Beta coefficients of these categories are .30, .27, .20 and .10 respectively (R^2 was .43). The categories SA and MC do not explain variance in the SI score.

7.3. Comparing changes detected by different SIP-versions

The SIP136 and SIP68 are meant to be evaluative instruments measuring health-related functional status. The 'SIP73' is studied as an alternative to the SIP68. Consequently, the changes detected by these three instruments should show a relatively high level of agreement. Hence, an other possibility to compare information

Table 7.5. Effect sizes of significant changes registered by SIP68, 'SIP73' and SIP136, correlation between changes detected by SIP68, 'SIP73' and SIP136

effect size = change detected/standard deviation of change detected

diagnosis	es SIP68	es 'SIP73'	es SIP136	r diff SIP136-SIP68	r diff SIP136-'SIP73'
rheumatoid arthritis	.73	.72	.68	.93	.92
ankylosing spondylitis	.46	.41	.38	.98	.98
spinalcord injury	.74	.77	.73	.92	.92
ankylosing spondylitis	.17	.15	.20	.88	.59
back/neck pain	.17	.16	.41	.59	.61
cancer	.56	.59	.72	.90	.92
Crohn's disease	.35	.33	.41	.94	.95

supplied by the SIP68 and the 'SIP73' is a comparison of the changes indicated by both instruments with each other and with changes detected by the SIP136. Longitudinal data was available from seven different diagnostic groups. This data stems from five different studies. As the research design in all five studies was different (different sizes of population, different periods between measurement moments), the data on changes presented in table 7.5 can not be aggregated across the studies. Table 7.5 presents the changes detected by the SIP68, the 'SIP73' and the SIP136 for every longitudinal population available, by means of effect sizes. A double line seperates the different studies. An effect size is a standardized and dimensionless representation of change registered by a measurement instrument. Several formulas exist to compute effect sizes. In this study it is calculated by dividing the difference in scores between two successive measurement moments (change detected) by the standard deviation of this difference. Table 7.5 shows that adding Social Isolation to the SIP68 only marginally influences the effect size of the SIP68. The correlations between changes found by the SIP68 and the 'SIP73' respectively with changes found by the SIP136 (last two columns of table 7.5) show that the 'SIP73' only marginally differs from the SIP68 as far as its relation to the responsiveness of the SIP136 is concerned.
At this point the preliminary conclusion is that no empirical evidence is found for the necessity to add the social isolation items to the SIP68.

7.4. Principal components analysis of the 'SIP73'

If Social Isolation is an independent and empirically relevant aspect of health-related functional status, a PCA of the 'SIP73' would distinguish it as a seperate factor on which mainly the five SI items load. Therefore, a principal components

analyses (PCA) was performed on the 'SIP73' (68 SIP68 items and 5 SI items). As it is supposed that SI might be a seperate category, the number of factors to be extracted was at first limited to 7 (the six SIP68 categories and SI). The solution found in this run did not confirm the presence of a seperate SI category. The first six factors extracted show a close resemblance with the six SIP68 factors. One SI-item loaded on the SIP68 factor Social Behavior, the other four loaded on the SIP68 factor Emotional Stability. The seventh factor existed out of four Mobility Control items and the item on trouble writing and typing from Psychic Autonomy and Communication. When the number of factors to be extracted was limited to six, almost the exact original SIP68 factors were found. The SI-items again loaded on the factors Social Behavior and Emotional Stability. Hence, it was concluded that Social Isolation cannot be distinguished as a separate category in our data.

7.5. Conclusion

The only interpretable factor found in the 40 items that were dropped from the original SIP136 because of their low factor loadings, was a five item factor interpreted as representing Social Isolation (SI). The SI score shows a relatively high correlation with the categories Social Behavior and Emotional Stability of the SIP68. Moreover, in a principal components analysis on the 'SIP73' (SIP68+SI items), SI items load on the factors that represent these SIP68 categories. This leads to the conclusion that the theoretically distinguishable category of behavior connected with social isolation is covered by the categories of Social Behavior and Emotional Stability in the SIP68. It might be that social isolation follows from behavior described in SIP68 categories SB and ES. Hence, SI is a result of SB and ES. Moreover, in theory it would probably not be easy to make a meaningful distinction between behavior expressing social isolation on the one hand and social behavior in general and behavior expressing the level of emotional stability on the other. Therefore it was concluded that the five SI-items do not supply information that is not yet incorporated in the SIP68. Moreover the concepts of Social Behavior and Emotional Stability cover the aspect of social isolation. Hence, adding SI-items to the exsisting SIP68 would be redundant and would not improve the validity or efficiency of the SIP68 as a measure of health-related functional status.

Appendix

Table 7.1. Six factor solution of PCA over 40 'SIPREST' items

	loading
1.	
I am not now using public transportation.	.58
I sit during much of the day.	.57
I get around only by using a walker, crutches, cane, wall, or furniture.	.54
I am not doing any of the maintenance or repair work that I would usually do in my home or yard.	.52
I walk only with help from someone.	.50
I have trouble getting shoes, socks, or stockings on.	.38
I am doing more inactive pastimes in place of my other usual activities.	.34
I hold on to something to move myself around in bed.	.32
2.	
I say how bad or useless I am, for example, thatI am a burden on others.	.58
I dress myself, but do so very slowly.	.53
I laugh or cry suddenly.	.48
I often express concern over what might be happening to my health.	.42
I am not doing any of my usual inactive recreation and pastimes, for example, watching TV, playing cards, reading.	.41
I get sudden frights.	.39
I feed myself but only by using specially prepared food or utensils	.39
I am only going to places with restrooms nearby.	.37
I move my hands or fingers with some limitation or difficulty	.36
I sometimes behave as if I were confused or disoriented inplace or time, for example, where I am, who is around, directions, what day it is.	
I am not doing the things I usually do to take care of my children or family.	
3.	
I sleep less at night, for example, wake up too early, don't fall asleep for a long time, awaken frequently	.52
I have more minor accidents, for example, drop things, trip and fall, bump into things.	.52
I act nervous or restless.	.49
I keep rubbing or holding areas of my body that hurt or are unconfortable.	.47
I do work around the house only for short periods of time or rest often.	.45
I change position frequently.	.35
I sleep or nap more during the day.	
4.	
I stay alone much of the time.	.57
I show less affection	.54
I am avoiding social visits from others	.46
I show less interest in other people's problems, for example, don't listen when they tell me about their problems, don't offer help.	.46
I am going out less to visit people.	.42
5.	
I don't write exept to sign my name.	.59
I am not doing any of my usual physical recreation or activities.	.53
I am eating special or different food, for example, soft food, bland diet, low-salt, low-fat, low-sugar.	.43
I am cutting down on some of my usual physical recreation or activities.	-.32
6.	
I am staying in bed more.	.66
I spend much of the day lying down in order to rest.	.57
I lie down more often during the day in order to rest.	.41
I am understood with difficulty.	-.34
I often lose control of my voice when I talk, for example, my voice gets louder or softer, trembles, changes unexpectedly.	-.32

Table 7.2. Percentage variance explained and eigenvalue of factors in SIP68 and 'SIPREST'

	SIP68		rest items	
	% variance	eigen value	% variance	eigen value
fact.1	15.3	11.1	12.8	5.1
fact.2	6.9	5.0	5.0	2.0
fact.3	4.7	3.3	4.3	1.7
fact.4	3.7	2.7	3.8	1.5
fact.5	3.1	2.2	3.6	1.4
fact.6	2.7	1.9	3.4	1.3
total	36.4		32.9	

Table 7.2A. Chronbach's α for factors found within the list of rest items

(n=2337)

	n items	Cronbach's α
total rest	40	.81
factor 1	8	.66
factor 2	11	.56
factor 3	7	.51
factor 4	5	.57
factor 5	4	.13
factor 6	5	.43

Literature

1. Bergner M, RA Bobbit, WB Carter, BS Gilson. The Sickness Impact Profile: development and final revision of a health status measure. Med.Care 1981, 21:787-805.
2. de Bruin AF, JPM Diederiks, LP de Witte, FCJ Stevens, H Philipsen: The development of a short generic version of the Sickness Impact Profile, J.Clin.Epid. 1994, vol. 47, no 4: 407-418.

Summary

In a number of research projects at the Institute for Rehabilitation Research (IRV) in Hoensbroek and at the University of Limburg, the Sickness Impact Profile (SIP) was used as a central measure of health status. The SIP is a generic measure of health related functional status. Questions concerning the SIP arose, that lead to the start of a project aimed at the following three main topics: the theoretical background of measuring health related functional status, psychometric characteristics of the SIP, and finally the development of a short version of the instrument. In this SIP-project secundairy analysis was perfomed on SIP-databases that were gathered in other research projects. This thesis is a report on the SIP-project.
With respect to the theoretical background of measuring health related functional status, based on sociological literature a model was developed that reflects the way in which health related behavioral changes are generated. from this model it can be derived that these behaviroal changes are not solely determined by bio-physiological factors. In most cases there is no a-priori and directly bio-physiologically determined set of 'necessary' behavioral changes. This leaves room for negotiations about the true, unavoidable level and type of behavioral changes resulting from a given state of health in bio-physiological terms. Based on their own interpretation of the health situation, all individuals involved (the patient, his social surroundings, and the relevant health professionals) will have expectations about the connected behavioral changes. In the negotiations these expectations and interpretation are confronted. Hence the resulting health related functional status is influenced by the level to which a patients social surroundings permits an individual to changes his behavior (willingness to take over tasks or pressure to change behavior). This process of observation, interpre-tation and negotiation is a circular one. Every percepted change in the health situation or in the intrepretation of the health status may lead to new negotiations and a redefinition of the health related functional status. Hence, health related functional status appeared to be a less objective concept that it is often thought to be.
The next step in the project was a literature review into the psychometric characteristics of the SIP. From this review it was concluded that the SIP is a reliable and valid generic measure of health. The major drawback of the instrument appeared to be its length. Hence it was decided to develop a short generic version of the instrument using available SIP data-bases. This resulted in the SIP68: a selection of 68 SIP items divided over six categories. This newly developed short instrument was sub-sequently tested for its validity and reliability as a measure of health related functional status. It appeared to be a valid and reliable alternative to the original Sickness Impact Profile. From the literature review it also occurred that relatively little is known about the ability of the SIP to detect relevant changes in functional status (responsiveness). In the last chapter of this thesis, therefore, based on literature

findings and secondary analysis of available databases, the responsivenss of the SIP and the SIP68 were studied and compared. As there is no generally accepted technique to quantify and judge responsiveness, it was not possible to reach a final judgement. However, literature findings indicated that the SIP generally shows score changes in the expected direction and in accordance with changes shown by related instruments. Secondary analysis of SIP-data bases learned that the responsiveness of the SIP68 does not significantly differ from the responsiveness of the original SIP. For both instruments it was found that changes are especially found in populations with relatively severe health deviations and in populations receiving intensive treatment in which psycho-social aspects of health deviations and functioning are attended to.

Samenvatting

In een aantal onderzoeksprojecten die gedaan werden op het IRV in Hoensbroek en aan de Rijksuniversiteit Limburg werd de Sickness Impact Profile (SIP) gebruikt als centrale maat voor gezondheid. De SIP is een generieke maat voor de gezondheid gerelateerde functionele toestand. Bij een aantal onderzoekers die de lijst gebruikten rezen vragen over de SIP, die ertoe geleid hebben dat het zogenaamde SIP-project gestart werd. Dit project richtte zich op drie hoofdpunten: de theoretische achtergronden van het meten van gezondheid gerelateerde functionele status, psychometrische kenmerken van de SIP, en tenslotte de mogelijkheden voor het ontwikkelen van een korte versie van het instrument. Bij de beantwoording van deze vragen kon gebruik gemaakt worden van de in het kader van andere onderzoeksprojecten verzamelde SIP-data bestanden. Dit proefschrift is een rapportage van het SIP-project.

In het kader van de theoretische achtergrond van het meten van functionele status werd, op basis van met name sociaal wetenschappelijke literatuur, een model ontwikkeld waarin wordt weergegeven langs welke weg gezondheid gerelateerde gedragsveranderingen tot stand komen. Uit dit model kan worden afgeleid dat niet slechts bio-physiologische factoren gedragsveranderingen bepalen. Omdat er in het algemeen geen a-priori en rechtstreeks uit de bio-physiologische toestand af te leiden 'noodzakelijk' patroon van gedragsveranderingen bekend is, is er ruimte voor onderhandeling over wat terechte of noodzakelijke gedragsaanpassingen zijn. De confrontatie van interpretaties van de gezondheidstoestand en de daarmee verbonden gedrags-verwachtingen van alle betrokkenen (de patiënt, de betrokken professionals en de sociale omgeving van de patiënt) beïnvloedt zodoende de uiteindelijke gezondheidgerelateerde functionele status. De mate waarin de sociale omgeving, zich verplicht voelt of bereid is taken over te nemen, of clementie te tonen, bepaalt hierdoor mede de functionele status. Anderzijds kan de interpretatie van de gezondheidsproblematiek ook leiden tot een (onterechte) beperking van de door de sociale omgeving verwachte of toegestane gedragsalternatieven.

Dit proces van interpretatie en onderhandeling is te beschouwen als een continu, circulair proces van observatie, interpretatie en feedback. Iedere gepercipieerde verandering in de gezondheids-situatie of de interpretatie daarvan kan leiden tot nieuwe onderhandelingen en een herdefinitie van het gepaste nivo van gezondheid gerelateerde functionele status. Al met al bleek de gezondheid gerelateerde functionele toestand een minder 'objectief' concept dan doorgaans wordt aangenomen.

De volgende stap in het project was een literatuur studie naar de psychometrische eigenschappen van de SIP. Hieruit kwam de SIP naar voren als een betrouwbare en valide generieke gezondheidsmaat is. Het grootste nadeel dat naar voren kwam was de relatief grote lengte van de lijst. Vandaar dat besloten werd op basis van

secundaire analyse over de beschikbare SIP-data bestanden een korte, maar eveneens generieke versie van het instrument te ontwikkelen. Dit resulteerde in een selectie van 68 items, verdeeld over zes categorieën: de SIP68. Deze verkorte lijst is vervolgens getoetst op validiteit en betrouwbaarheid. Het bleek dat de SIP68 een generiek, valide en betrouwbaar alternatief is voor de originele Sickness Impact Profile.

Een tweede punt dat uit de literatuur studie naar de SIP naar bleek was dat er vrij weinig bekend is over de capaciteiten van de lijst om relevante veranderingen in de functionele status weer te geven (de responsibiliteit). Het laatste hoofdstuk van dit proefschrift is dan ook gewijd aan dit onderwerp. In dit hoofdstuk wordt, op basis van literatuur onderzoek en met behulp van secundaire analyse, de responsibiliteit van de SIP en van de SIP68 onderzocht en vergeleken. Het bleek niet mogelijk absolute uitspraken te doen over de responsibiliteit. Er is namelijk geen algemeen aanvaarde techniek of norm voor het beoordelen van dit kenmerk. Op basis van literatuurgege-vens werd geconcludeerd dat de SIP veranderingen aangeeft in de verwachte richting en in overeenstemming met veranderingen aangegeven door verwante instrumenten. Uit de secundaire analyses bleek dat de veranderingsgevoeligheid van de SIP68 niet significant verschilt van die van het originele instrument. Voor beide instrumenten bleek dat met name veranderingen werden gevonden in populaties die relatief zwaar aangedaan waren en bij intensieve behandeling waarbij met name de psycho-sociale aspecten van de ziekte of het functioneren betrokken werden in de behandeling.

Curriculum Vitae

Op 27 juli 1960 werd ik geboren in Groningen. De Lagere School begon ik in Dieren en maakte ik af in Enschede, waar ik in 1979 mijn VWO diploma behaalde. Hierna ging het wat minder 'rechtdoor': een half jaar studeerde ik Bestuurskunde in Enschede, gevolgd door een half jaar werken als 'intern transporteur' en als magazijnchef in een revalidatiecentrum.
Vanaf 1980 studeerde ik Logopedie aan de toenmalige Revalidatie-academie te Hoensbroek. Voordat ik hier het diploma behaalde, was ik al begonnen met de studie Gezondheidswetenschappen aan de Rijksuniversiteit Limburg in Maastricht. Hier koos ik de richting Beleid en Beheer van Gezondheidszorgvoorzieningen, waarin ik in 1988 afstudeerde.
Tijdens de studie in Maastricht begon ik bij de Maastrichtse Studententoneel-vereniging 'Alles is Drama' aan een nog voortdurende parallel-carrière als amateur-acteur.

Na mijn afstuderen was ik korte tijd projectmedewerker in het Academisch Ziekenhuis Utrecht op de afdeling Informatie Organisatie. In 1989 begon ik mijn werkzaamheden aan het SIP-project bij de Vakgroep Medische Sociologie van de Rijksuniversiteit Limburg; aanvankelijk als AIO, later als universitair docent. Naast onderzoekswerkzaamheden in het kader van mijn proefschrift hield ik mij bezig met het geven van onderwijs aan de Faculteit der Gezondheidswetenschappen (met name in de methoden en technieken van wetenschappelijk onderzoek) en was ik gedurende vrijwel mijn gehele aanstellingsduur lid van het dagelijks bestuur van de vakgroep. De laatste periode dat ik bij Medische Sociologie werkte, hoorde ook een deel van het beheer van de vakgroep tot mijn takenpakket.

Sedert 15 juli 1995 tenslotte, ben ik werkzaam als hoofd Beheer van de afdeling Revalidatie van de Ziekenhuis De Wever & Gregorius met locaties te Heerlen en Brunssum en het St. Jozefziekenhuis te Kerkrade.

In dit curriculum mag niet onvermeld blijven dat ik in 1993 gehuwd ben met Ineke Kortland en sinds mei 1995 vader ben van Daan.

Dankwoord

Promoveren doe je meestal alleen, het onderzoek waarop je promoveert doe je nooit in je eentje. Dit is ongeveer het clichébegin van een dankwoord en dat is niet voor niets. Ook aan het welslagen van het SIP-project hebben een groot aantal mensen een bijdrage geleverd die ik in dit dankwoord wil noemen.
Allereerst uiteraard de SIP-project-begeleidingsgroep. Mijn promotor professor Hans Philipsen die de grote lijn in het project heeft bewaakt en op cruciale momenten helderheid wist te brengen in complexe situaties. Jos Diederiks, mijn co-promotor waarmee ik met name in het laatste deel van het project intensief en plezierig heb samengewerkt. Ook aan de gesprekken over onderzoek in de revalidatie en aan onze gemeenschappelijke belangstelling voor cabaretteksten, denk ik met plezier terug. Luc de Witte, tijdens mijn studie kwam ik via jouw promotieproject met de SIP in aanraking en je ziet waar dat toe geleid heeft. Je consciëntieuse en efficiënte wijze van begeleiden heeft, behalve een afstudeerscriptie, nu ook nog een proefschrift tot gevolg gehad. Fred Stevens was nauw betrokken bij het gehele project en begeleidde met name mijn eerste wankele aio-schreden op het bochtige onderzoekspad.

Gezien het feit dat het SIP-project voor een groot deel gebaseerd is op secundaire analyse wil ik zeker ook de dataleveranciers bedanken: Annemie Courtens, Eddy van Doorslaer, Marianne Geerts, Alita Hidding, Miriam Janssen, Jolande Keijzers, Bart Koes, Jacqueline Peeters, Marcel Post, Sjoerd Terpstra, Anne Visser en Luc de Witte. Dank voor de vanzelfsprekendheid waarmee ik over jullie data mocht beschikken. Het koppelen van jullie bestanden heeft mij geleerd dat er vele manieren zijn om een databestand op te bouwen.
Twee mede-auteurs mogen zeker niet onvermeld blijven. Marcel Post, doordat ik jouw dwarslaesiedata mocht gebruiken en mee kon schrijven aan ons artikel, heb ik een hoofdstuk aan mijn proefschrift toe kunnen voegen. Dank voor deze prettige samenwerking. Marja Buys heeft een cruciale bijdrage geleverd aan het welslagen van het onderzoeks-subproject in Venlo en bij het schrijven van het verslag hiervan. Ook dit heeft een publicatie, alsmede een hoofdstuk in dit proefschrift opgeleverd. Dank hiervoor.

Het corrigeren van mijn 'Engels' is voor het grootste deel gedaan door Thomas Swaak. Aan hem is het te danken dat ik nooit om taalkundige redenen een manuscript teruggestuurd heb gekregen. Het manuscript van het eerste hoofdstuk is door Bob Wilkinson roodgekleurd met taalkundige verbeteringen, en daar is het flink van opgeknapt. De finishing touch voor wat betreft de vormgeving van dit proefschrift is van de hand en muis van Hein Berendsen. Hierdoor kreeg ook het oog nog wat het wil.

Tenslotte is er nog een aantal mensen dat een minder directe, maar zeker niet minder relevante bijdrage heeft geleverd aan het SIP-project. Allereerst de collega's van de Vakgroep Medische Sociologie. Het is mede aan hen te danken dat ik met veel plezier zes jaar 'in het onderzoek' heb doorgebracht. Met name wil ik in dit verband mijn beide paranimfen noemen: Theo en Annemie, onze trio-lunches en overige binnen- en buiten-werktijdse activiteiten behoren tot de prettigste herinneringen aan mijn tijd bij de RL. Gelukkig worden die aktiviteiten gecontinueerd, nu ik 'in het veld' werk.

Ook het einde van een dankwoord is bijna een cliché, maar ook dat is niet voor niets: ik dank mijn ouders die mij de gelegenheid gaven te studeren wat ik wilde. En tenslotte: Ineke bedankt voor jouw bijdrage die gelukkig niet tot dit project beperkt blijft.